第 5 章　CMOS 反相器

■ 反相器完整性、性能和能量指标的定量分析
■ 反相器设计的优化

5.1　引言

反相器确实是所有数字设计的核心。一旦清楚理解了它的工作和性质，设计诸如逻辑门、加法器、乘法器和微处理器等比较复杂的结构就大大地简化了。这些复杂电路的电气特性几乎完全可以由反相器中得到的结果推断出来。反相器的分析可以延伸来解释比较复杂的门（如 NAND、NOR 或 XOR）的特性，它们又可以形成建筑块来构成如乘法器和处理器这样的模块。

本章将集中讨论一种具体的反相器门——静态 CMOS 反相器。这当然是目前最普遍的反相器，因此值得给予特别关注。我们对门的分析着眼于几个不同的设计指标，这些指标在第 1 章中已经予以概括：

- 成本：用复杂性和面积来表示
- 完整性和稳定性：用静态（即稳态）特性来表示
- 性能：由动态（即瞬态）响应决定
- 能量效率：由能耗和功耗决定

利用这一分析，我们建立了这个门的模型并确定了它的设计参数。我们还建立了选择这些参数值的方法，以使最终的设计能满足所希望的技术要求。虽然这些参数中的每一个对一种给定的工艺都能很容易地定量化，但我们还要讨论它们将如何受到工艺缩小的影响。

尽管本章仅集中在 CMOS 反相器，但在下一章中我们将看到同样的方法也适用于其他门的拓扑结构。

5.2　静态 CMOS 反相器——直观综述

图 5.1 显示了一个静态 CMOS 反相器的电路图。借助我们在第 3 章介绍的 MOS 晶体管的简单开关模型（见图3.26）可以很容易理解它的工作原理。晶体管只不过是一个具有无限关断电阻（当 $|V_{GS}| < |V_T|$ 时）和有限导通电阻（当 $|V_{GS}| > |V_T|$ 时）的开关。这产生了如下对反相器的解释。当 V_{in} 为高并等于 V_{DD} 时，NMOS 管导通而 PMOS 管截止。由此得到了图 5.2（a）的等效电路。此时在 V_{out} 和接地节点之间存在一个直接通路，形成一个稳态值 0 V。相反，当输入电压为低（0 V）时，NMOS 和 PMOS 管分别关断和导通。由图5.2（b）的等效电路可知在 V_{DD} 和 V_{out} 之间存在一条通路，产生了一个高电平输出电压。显然这个门具有反相器的功能。

从这个开关级的角度可以推导出静态 CMOS 的许多其他重要特性：

- 输出高电平和低电平分别为 V_{DD} 和 GND。换言之，电压摆幅等于电源电压。因此噪声容限

很大。

- 逻辑电平与器件的相对尺寸无关，所以晶体管可以采用最小尺寸。具有这一特点的门称为无比逻辑。它不同于有比逻辑，在有比逻辑中逻辑电平是由组成逻辑的晶体管的相对尺寸来决定的。
- 稳态时在输出和 V_{DD} 或 GND 之间总存在一条具有有限电阻的通路。因此一个设计良好的CMOS反相器具有低输出阻抗，这使它对噪声和干扰不敏感。输出电阻的典型值在kΩ的范围内。
- CMOS反相器的输入电阻极高，因为一个MOS管的栅实际上是一个完全的绝缘体，因此不取任何dc（直流）输入电流。由于反相器的输入节点只连到晶体管的栅上，所以稳态输入电流几乎为零。理论上，单个反相器可以驱动无穷多个门（或者说具有无穷大的扇出）而仍能正确工作，但我们很快会看到增加扇出也会增加传播延时。尽管扇出不会对稳态特性有任何影响，但它使瞬态响应变差。
- 在稳态工作情况下电源线和地线之间没有直接的通路（即此时输入和输出保持不变）。没有电流存在（忽略漏电流）意味着该门并不消耗任何静态功率。

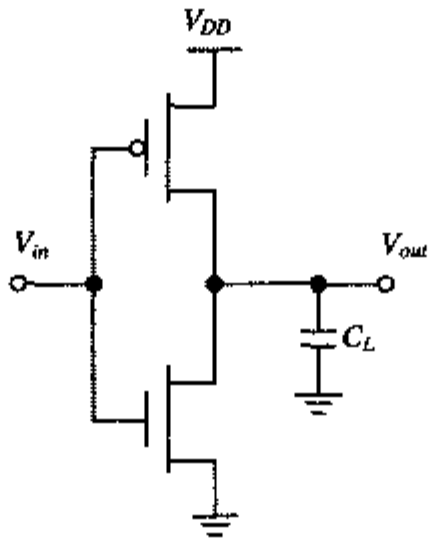


图 5.1　静态 CMOS 反相器。V_{DD} 代表电源电压

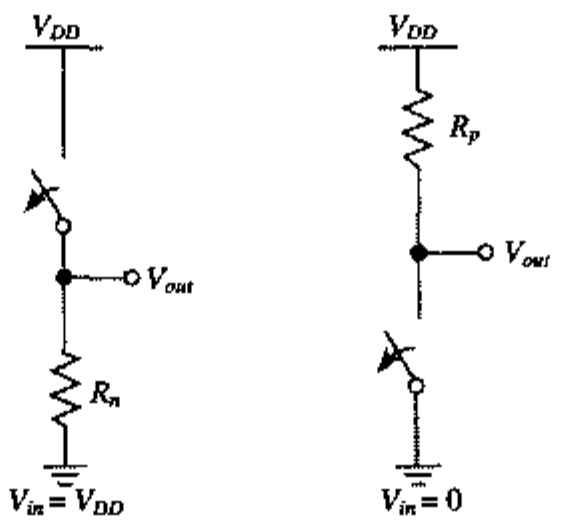


图 5.2　CMOS 反相器的开关模型

注：上面的观察虽然看起来很明显，但却是非常重要的，它是目前数字技术选择CMOS的主要原因之一。但在20世纪70年代以及20世纪80年代早期的情况则大不相同。所有早期的微处理器（如Intel 4004）都是只用NMOS工艺实现的。在这一工艺中由于缺少互补器件（如NMOS和PMOS管），所以很不容易使反相器具有零静态功耗。所产生的静态功耗严格限制了单片上能集成的逻辑门的最多数目，因此到20世纪80年代，当工艺技术缩小到允许更高集成密度时不得不转向CMOS。

电压传输特性（VTC）的性质和形状可以通过图解法迭加NMOS和PMOS器件的电流特性来得到。这样的一个图形结构通常称为负载曲线图。它要求把NMOS和PMOS器件的*I-V*曲线转换到一组公共坐标上。我们以输入电压V_{in}、输出电压V_{out}和NMOS漏电流I_{DN}作为选择的变量，于是PMOS的*I-V*关系就可以通过以下关系转换到这一变量空间中（下标*n*和*p*分别表示NMOS和PMOS器件）：

$$\begin{aligned} I_{DSp} &= -I_{DSn} \\ V_{GSn} &= V_{in};\quad V_{GSp} = V_{in} - V_{DD} \\ V_{DSn} &= V_{out};\quad V_{DSp} = V_{out} - V_{DD} \end{aligned} \tag{5.1}$$

PMOS器件的负载曲线可以通过对*x*轴求镜像并向右平移V_{DD}来得到。这一过程概括在图5.3中，它显示了将原先的PMOS *I-V*曲线调整至公共坐标系V_{in}、V_{out}和I_{Dn}的一系列步骤。

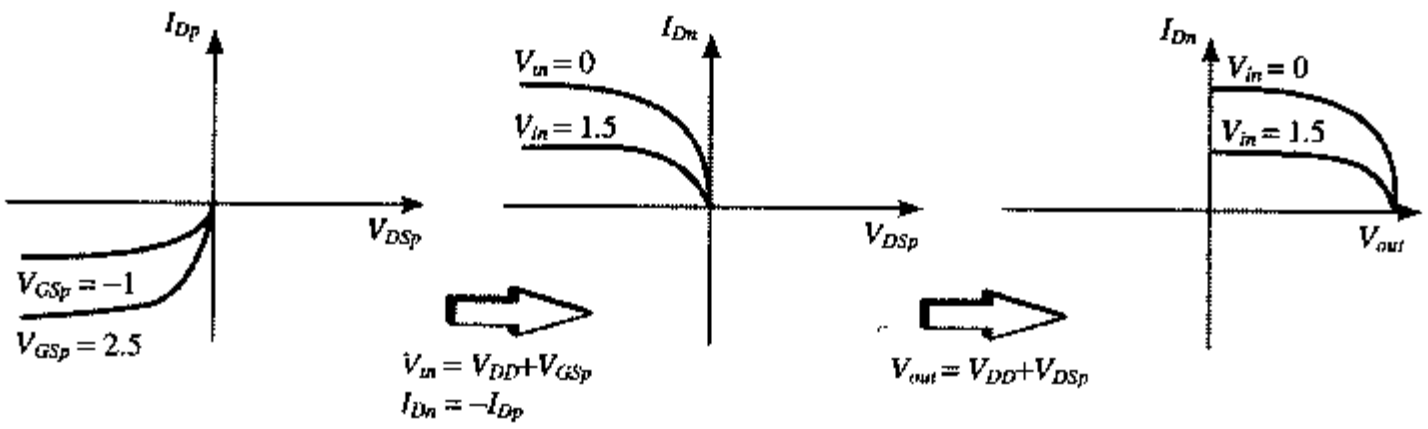


图5.3 将PMOS *I-V*特性转换至公共坐标系（假设$V_{DD}=2.5$ V）

所得到的负载线画在图5.4中。为使一个dc工作点成立，通过NMOS和PMOS器件的电流必须相等。用图解法时这意味着dc工作点必须处在两条相应负载线的交点上。图上标记了许多这样的点（对$V_{in}=0, 0.5, 1, 1.5, 2$和2.5 V）。可以看到，所有的工作点不是在高输出电平就是在低输出电平上。因此反相器的VTC显示出具有非常窄的过渡区。这是由于在开关过渡期间的高增益造成的，此时NMOS和PMOS同时导通且处于饱和状态。在这一工作区，输入电压的一个很小变化就会引起输出的很大改变。所有这些观察结果都可以用VTC形式显示在图5.5中。

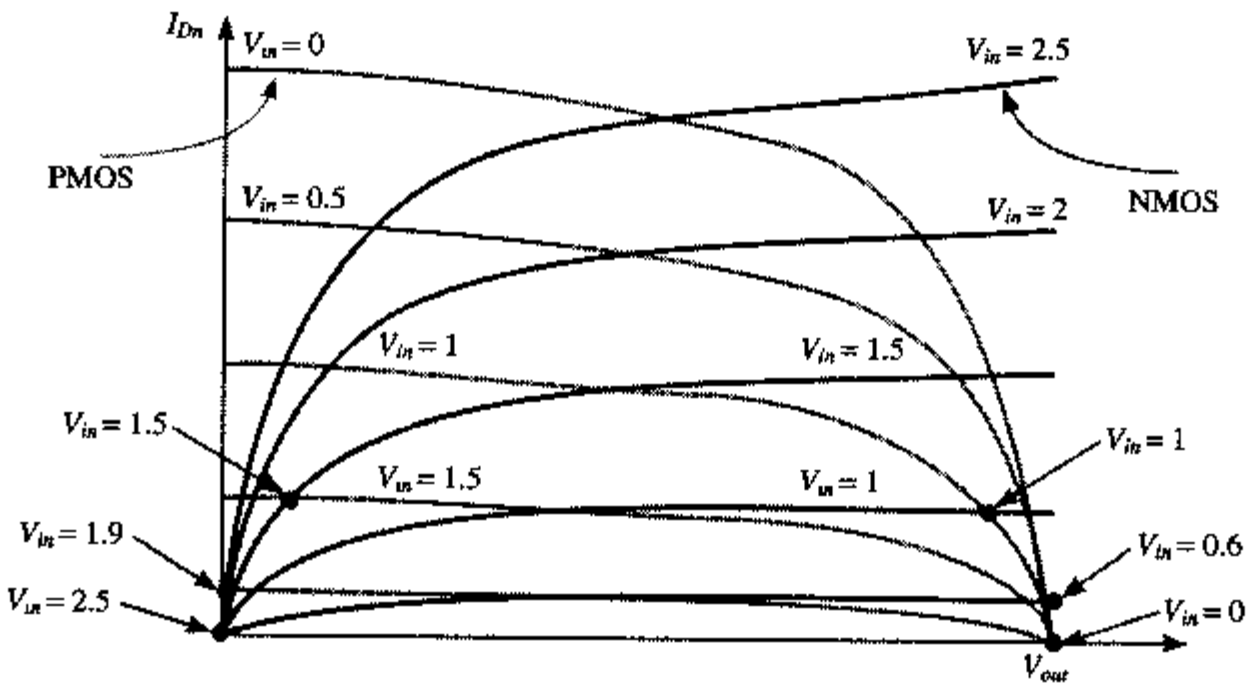


图5.4 静态CMOS反相器中NMOS和PMOS管的负载曲线（$V_{DD}=2.5$ V）。圆点代表各种输入电压下的dc（直流）工作点

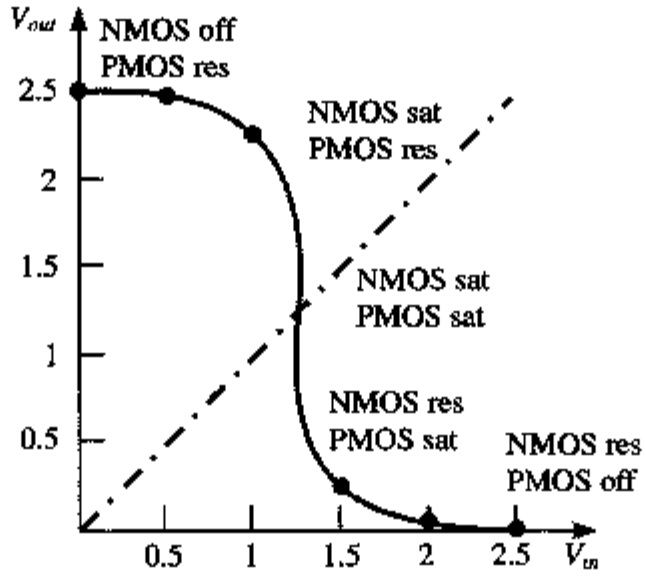


图 5.5　由图5.4（$V_{DD}=2.5$ V）推导出的CMOS反相器的VTC。对于每一个工作区都标注了晶体管的工作模式——off（截止）、res（电阻模式）或sat（饱和）

在对CMOS反相器的工作进行细节分析之前，最好先对这个门的瞬态特性进行定性分析。这一响应主要由门的输出电容 C_L 决定，它包括NMOS和PMOS晶体管的漏扩散电容、连线电容以及扇出门的输入电容。暂且假设晶体管的切换是瞬时发生的，我们可以再次利用简化的开关模型来得到一个近似的瞬态响应的概念。首先考虑由低至高的过渡［见图5.6（a）］。门的响应时间是由通过电阻 R_p 充电电容 C_L 所需要的时间决定的。在例4.5中，我们了解到这样一个电路的传播延时正比于时间常数 R_pC_L。因此，一个快速门的设计是通过减小输出电容或者减小晶体管的导通电阻实现的。后者可以通过加大器件的 W/L 比来到。同样的考虑对于高至低的过渡也成立［见图5.6（b）］，这一过渡取决于时间常数 R_nC_L。读者应当注意，NMOS和PMOS晶体管的导通电阻并不是常数，而是晶体管两端电压的非线性函数。这使确切决定传播延时变得比较复杂。在5.4节中我们将对如何分析和优化静态CMOS反相器的性能进行深入的分析。

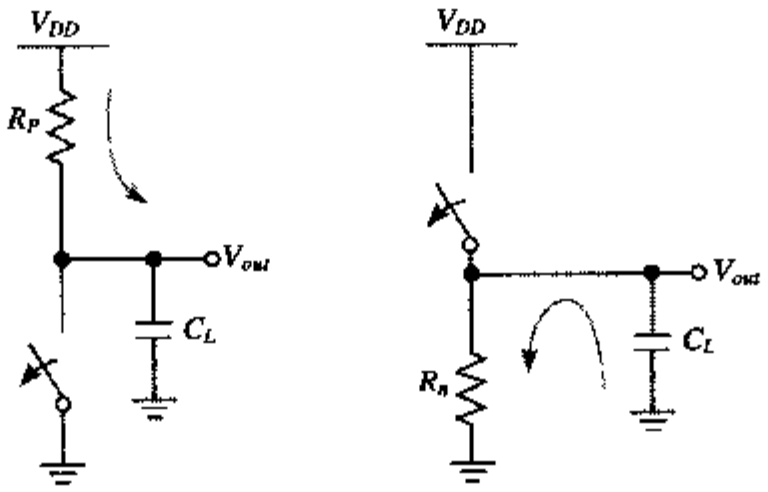


图 5.6　静态 CMOS 反相器动态特性的开关模型

5.3　CMOS 反相器稳定性的评估——静态特性

在以上的定性讨论中我们简述了静态CMOS反相器电压传输特性的概貌，并推导了 V_{OH} 和 V_{OL} 的值——它们分别估计为 V_{DD} 和 GND。接下来要确定 V_M、V_{IH} 和 V_{IL} 以及噪声容限的精确值。

5.3.1　开关阈值

开关阈值 V_M 定义为 $V_{in}=V_{out}$ 的点，其值可以用图解法由VTC与直线 $V_{in}=V_{out}$ 的交点求得（见

图5.5）。在这一区域由于$V_{DS}=V_{GS}$，PMOS和NMOS总是饱和的。使通过两个晶体管的电流相等就可以得到V_M的解析表达式。我们求解的情形是电源电压足够高，所以这两个器件可以被假设为都处于速度饱和（即$V_{DSAT}<V_M-V_T$）。同时我们忽略沟长调制效应，于是有：

$$k_n V_{DSATn}\left(V_M - V_{Tn} - \frac{V_{DSATn}}{2}\right) + k_p V_{DSATp}\left(V_M - V_{DD} - V_{Tp} - \frac{V_{DSATp}}{2}\right) = 0 \qquad (5.2)$$

求解V_M得到：

$$V_M = \frac{\left(V_{Tn} + \frac{V_{DSATn}}{2}\right) + r\left(V_{DD} + V_{Tp} + \frac{V_{DSATp}}{2}\right)}{1+r} \qquad \text{其中 } r = \frac{k_p V_{DSATp}}{k_n V_{DSATn}} = \frac{\upsilon_{satp} W_p}{\upsilon_{satn} W_n} \qquad (5.3)$$

这里，假设PMOS和NMOS管的栅氧厚度相同。当V_{DD}值较大时（与晶体管阈值电压及饱和电压相比），公式（5.3）可以简化为：

$$V_M \approx \frac{rV_{DD}}{1+r} \qquad (5.4)$$

公式（5.4）表明开关阈值取决于比值r，它是PMOS和NMOS管相对驱动强度的比。一般希望V_M处在电压摆幅的中点（即$V_{DD}/2$处）附近，因为这可以使低电平噪声容限和高电平噪声容限具有相近的值。为此要求r接近1，这相当于使PMOS器件的尺寸为$(W/L)_p=(W/L)_n\times(V_{DSATn}k'_n)/(V_{DSATp}k'_p)$。为使$V_M$上移，要求$r$的值较大。反之提高NMOS的强度将使开关阈值趋近GND。

由公式（5.2）可以推导出使开关阈值等于所希望的值V_M时所要求的PMOS和NMOS管的尺寸：

$$\frac{(W/L)_p}{(W/L)_n} = \frac{k'_n V_{DSATn}(V_M - V_{Tn} - V_{DSATn}/2)}{k'_p V_{DSATp}(V_{DD} - V_M + V_{Tp} + V_{DSATp}/2)} \qquad (5.5)$$

使用这一表达式时要确认，对于所选择的工作点，这两个器件都处于速度饱和的假设仍然成立。

思考题 5.1 针对长沟道器件或低电源电压的反相器开关阈值

上面的表达式是在假设晶体管达到速度饱和的前提下推导的。当PMOS和NMOS为长沟道器件或电源电压较低时不发生速度饱和（即$V_M-V_T<V_{DSAT}$）。此时V_M的计算公式如下：

$$V_M = \frac{V_{Tn} + r(V_{DD} + V_{Tp})}{1+r} \qquad \text{其中 } r = \sqrt{\frac{-k_p}{k_n}} \qquad (5.6)$$

试推导该公式。

设计技术 使噪声容限最大

在设计静态CMOS电路时，若希望使噪声容限最大并得到对称的特性，建议使PMOS部分比

NMOS 部分宽以均衡晶体管的驱动强度。所要求的宽度比见公式（5.5）。 ■

例 5.1　CMOS 反相器的开关阈值

现在推导出PMOS和NMOS晶体管的尺寸，使以通用0.25 μm CMOS工艺实现的一个CMOS反相器的开关阈值处在电源电压的中点处。我们所用的工艺参数在例 3.7 中提供，并假设电源电压为2.5 V。最小尺寸器件的宽长比为1.5。利用公式（5.5），我们得到：

$$\frac{(W/L)_p}{(W/L)_n} = \frac{115\times10^{-6}}{30\times10^{-6}} \times \frac{0.63}{1.0} \times \frac{(1.25-0.43-0.63/2)}{(1.25-0.4-1.0/2)} = 3.5$$

图5.7画出了通过电路模拟得到的开关阈值与PMOS对NMOS比的关系。PMOS对NMOS比为3.4时由模拟得到开关阈值为1.25 V，这与从公式（5.5）中预见的值一致。

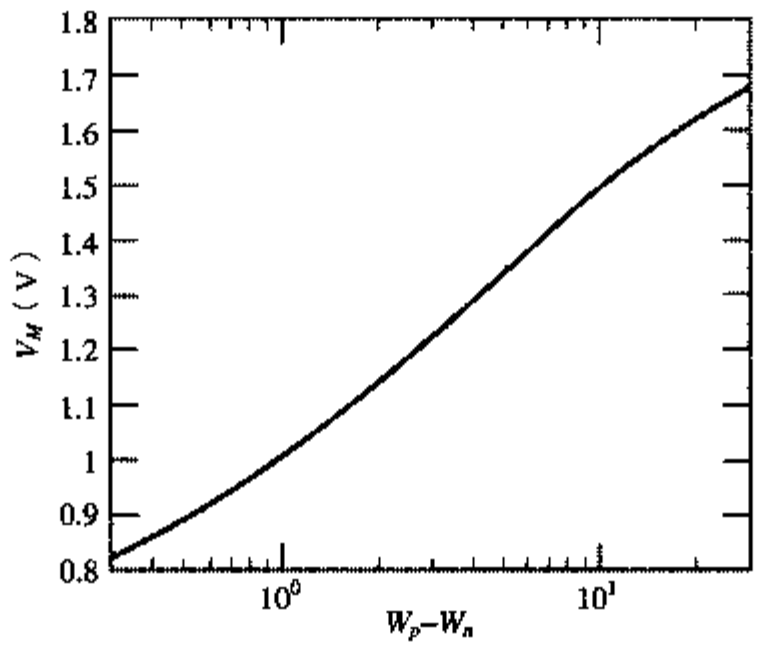
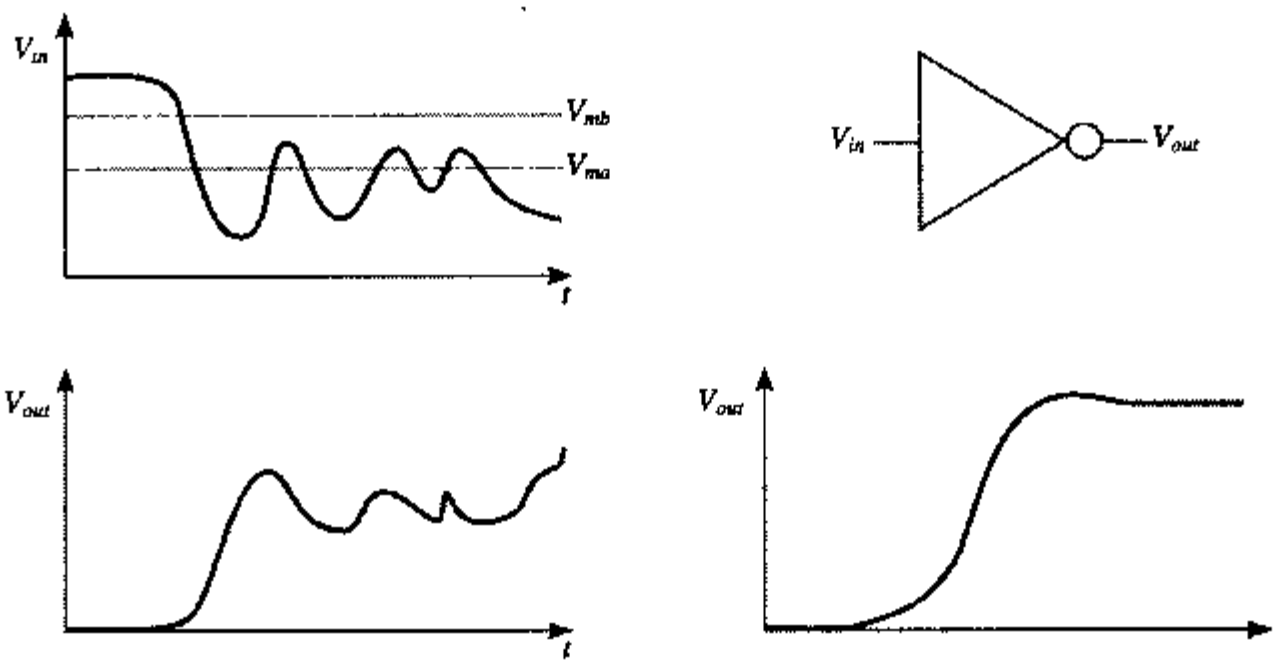


图 5.7　模拟得到的反相器开关阈值与 PMOS 对 NMOS 尺寸比的关系（0.25 μm CMOS，V_{DD}= 2.5 V）

通过对图5.7曲线的分析可以得到一些有意义的结果：

1. V_M 对于器件比值的变化相对来说是不敏感的。这意味着比值的较小变化（如使它为 3 或 2.5）并不会对传输特性产生多大的影响，因此在工业设计中使 PMOS 管的宽度小于完全对称时所要求的值是可以接受的。在前面的例子中将比值设为3，2.5和2，产生的开关阈值分别为1.22 V，1.18 V 和1.13 V。
2. 改变 W_p 对 W_n 比值的影响是使VTC的过渡区平移。增加PMOS或NMOS宽度使 V_M 分别移向 V_{DD} 或 GND。这一特性非常有用，因为不对称的传输特性实际上在某些设计中是所希望的。图5.8的例子即显示了这一点。输入信号 V_{in} 的零值受噪声严重干扰。若使这一信号通过一个对称的反相器，则会产生错误的输出值［见图5.8（a）］。这可以通过提高反相器的阈值来解决，结果得到一个正确的响应［见图5.8（b）］。在本书的后面我们还会看到其他一些电路例子需要反相器具有不对称的开关阈值。然而，要较大程度地改变开关阈值并不容易，特别是在电源电压与晶体管阈值的比相对较小时尤为如此（在我们特定的例子中为2.5/0.4 = 6）。要使阈值变为1.5 V需要管子宽长比为 11，而且要进一步加大阈值则代价很高，因而难以实现。注意，图5.7是以半对数坐标形式画出的。



（a）标准反相器的响应　　（b）改变阈值后的反相器的响应

图 5.8　改变反相器的阈值可以提高电路的可靠性

5.3.2　噪声容限

根据定义，V_{IH} 和 V_{IL} 是 $\frac{dV_{out}}{dV_{in}} = -1$ 时反相器的工作点。用模拟电路设计者的术语来说，它们是由反相器构成的放大器的增益 g 等于−1 时的点。虽然确实可以推导出 V_{IH} 和 V_{IL} 的解析表达式，但它们往往使用不便并且对深刻理解什么参数对确定噪声容限有用几乎没什么帮助。

比较简单的方法是对 VTC 采用逐段线性近似，如图 5.9 所示。过渡区可以近似为一段直线，其增益等于在开关阈值 V_M 处的增益 g。它与 V_{OH} 及 V_{OL} 线的交点用来定义 V_{IH} 和 V_{IL} 点。由此引起的误差很小并完全处在初步设计所要求的范围内。由这一方法可以得到过渡区的宽度 $V_{IH}-V_{IL}$，V_{IH}，V_{IL} 以及噪声容限 NM_H 和 NM_L 的表达式如下：

$$V_{IH}-V_{IL} = \frac{(V_{OH}-V_{OL})}{g} = \frac{-V_{DD}}{g}$$

$$V_{IH} = V_M - \frac{V_M}{g} \qquad V_{IL} = V_M + \frac{V_{DD}-V_M}{g} \tag{5.7}$$

$$NM_H = V_{DD}-V_{IH} \qquad NM_L = V_{IL}$$

这些表达式更清楚地表明在过渡区有较高的增益是我们所希望的。在增益为无穷大的极端情形下，噪声容限 NM_H 和 NM_L 分别简化为 $V_{OH}-V_M$ 和 V_M-V_{OL}，它们跨越了整个电压摆幅。

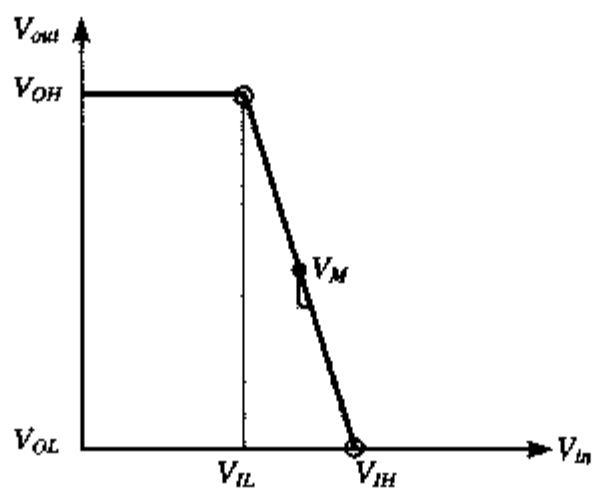


图 5.9　对 VTC 进行逐段线性近似简化了 V_{IL} 和 V_{IH} 的推导

接下来要决定的是静态CMOS反相器的中点增益。我们再次假设PMOS和NMOS都处在速度饱和。由图 5.4 可以清楚地看到，在饱和区增益与电流的斜率关系很大。因此在这一分析中不能忽略沟长调制系数，否则会导致增益无穷大。现在可以通过在电流公式（5.8）中对V_{in}求导数来推导出增益，公式（5.8）在开关阈值附近成立：

$$k_n V_{DSATn}\left(V_{in}-V_{Tn}-\frac{V_{DSATn}}{2}\right)(1+\lambda_n V_{out})+ k_p V_{DSATp}\left(V_{in}-V_{DD}-V_{Tp}-\frac{V_{DSATp}}{2}\right)(1+\lambda_p V_{out}-\lambda_p V_{DD}) = 0 \tag{5.8}$$

求导并求解 dV_{out}/dV_{in} 得到：

$$\frac{dV_{out}}{dV_{in}} = -\frac{k_n V_{DSATn}(1+\lambda_n V_{out})+k_p V_{DSATp}(1+\lambda_p V_{out}-\lambda_p V_{DD})}{\lambda_n k_n V_{DSATn}(V_{in}-V_{Tn}-V_{DSATn}/2)+\lambda_p k_p V_{DSATp}(V_{in}-V_{DD}-V_{Tp}-V_{DSATp}/2)} \tag{5.9}$$

忽略某些二次项并令 $V_{in}=V_M$，将得到如下的增益表达式：

$$g = -\frac{1}{I_D(V_M)}\frac{k_n V_{DSATn}+k_p V_{DSATp}}{\lambda_n-\lambda_p} \approx \frac{1+r}{(V_M-V_{Tn}-V_{DSATn}/2)(\lambda_n-\lambda_p)} \tag{5.10}$$

其中，$I_D(V_M)$是$V_{in}=V_M$时流过反相器的电流。这一增益几乎完全取决于工艺参数，特别是沟长调制。设计者通过选择电源电压及晶体管尺寸只能对它产生很小的影响。

例 5.2　CMOS 反相器的电压传输特性和噪声容限

假设设计一个通用0.25 μm CMOS工艺的反相器，PMOS对NMOS的比为3.4，其中NMOS晶体管的最小尺寸为（$W=0.375$ μm，$L=0.25$ μm，即 $W/L=1.5$）。我们首先计算在V_M(= 1.25 V)处的增益：

$$I_D(V_M) = 1.5\times 115\times 10^{-6}\times 0.63\times(1.25-0.43-0.63/2)\times(1+0.06\times 1.25) = 59\times 10^{-6}\ \text{A}$$

$$g = -\frac{1}{59\times 10^{-6}}\frac{1.5\times 115\times 10^{-6}\times 0.63+1.5\times 3.4\times 30\times 10^{-6}\times 1.0}{0.06+0.1} = -27.5 \tag{5.10a}$$

由此得到如下 V_{IL}，V_{IH}，NM_L，NM_H 的值：

$$V_{IL}=1.2\ \text{V},\ V_{IH}=1.3\ \text{V},\ NM_L=NM_H=1.2$$

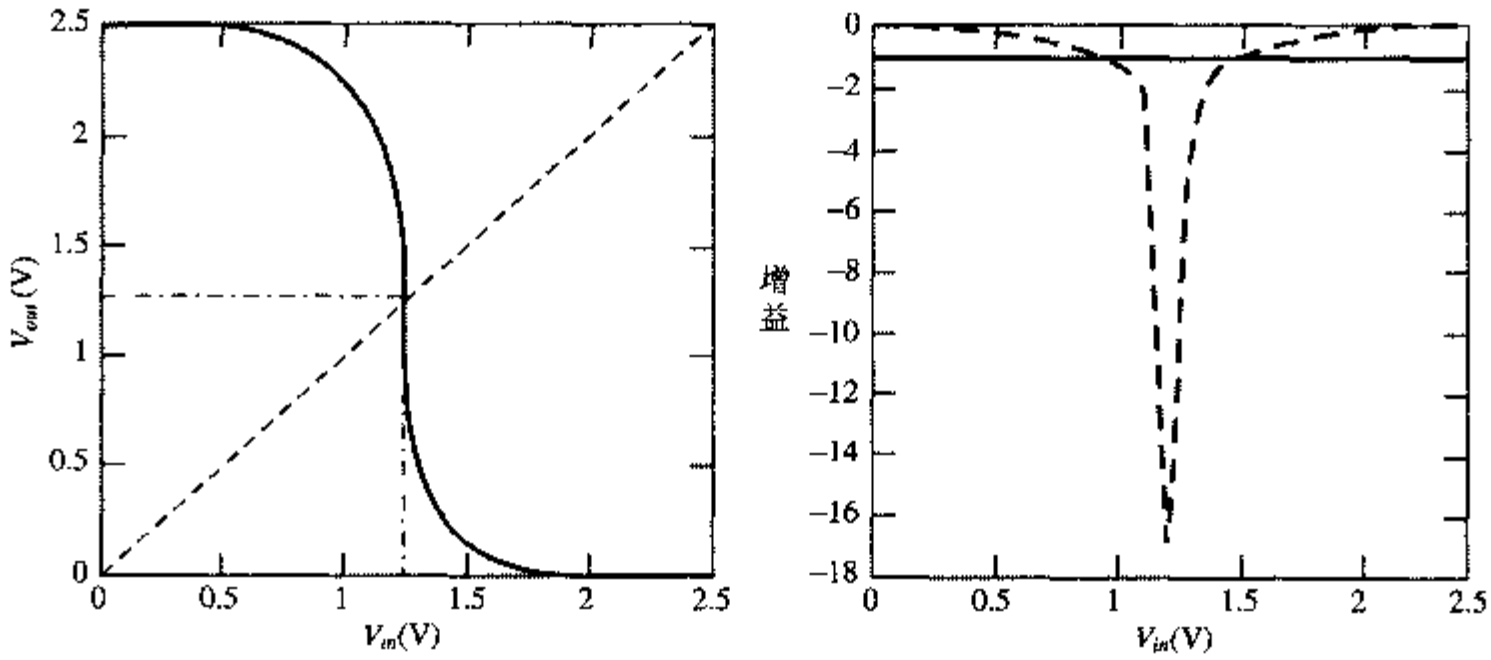
图 5.10 画出了模拟得到的反相器的VTC以及它的导数（即增益）。可见其非常接近理想特性。V_{IL}和V_{IH}的确切值分别为1.03 V和1.45 V，使噪声容限为1.03 V和1.05 V。由于以下两个原因这两个值低于预测的值：

- 公式（5.10）过高估计了增益。由图 5.10（b）可知最大增益（在V_M处）仅等于 17。这一减

少的增益使 V_{IL} 和 V_{IH} 的值分别为1.17 V 和1.33 V[①]。

- 最大的偏离是由于对VTC的逐段线性近似造成的。这对于实际的噪声容限而言是过于乐观的。然而这里得到的公式对于一阶估计以及作为识别相关参数及其影响的工具来说是绝对有用的。

最后，在这个例子中我们也可以通过模拟提取反相器在低电平和高电平状态时的输出电阻。可以看到它的最小值分别为 2.4 kΩ 和 3.3 kΩ。输出电阻很好地衡量了该门对于在输出端引起的噪声的敏感程度，我们希望它尽可能地小。



（a）电压传输特性　　（b）电压增益

图 5.10　模拟得到的 CMOS 反相器（0.25 μm，V_{DD} = 2.5 V）的电压传输特性和电压增益

注：出人意料的是，静态CMOS反相器也可以作为一个模拟放大器，因为它在过渡区具有相当高的增益。然而这一区域非常窄，正如图5.10（b）清楚显示的那样。而且它作为放大器所具有的其他一些性质也很差，如电源噪声抑制等。这一观察还可以用来说明模拟和数字设计之间的一个主要差别。模拟设计者会把这一放大器偏置在过渡区的中点以得到最大的线性度，而数字设计者将使该器件工作在极端的非线性区域，从而得到定义明确和分离得很好的高、低电平信号。

思考题 5.2　长沟道器件反相器的噪声容限

假设PMOS和NMOS是长沟道器件（或电源电压为低电压），因而不发生速度饱和，推导出增益和噪声容限的表达式。

5.3.3　再谈稳定性

器件参数变化

虽然我们设计一个门时都是针对通常的工作情况和典型的器件参数，但我们总是应当记住实际的工作温度会在一个很大的范围内变化，而且制造后的器件参数将偏离在设计优化过程中所采用的典型值。所幸的是，静态 CMOS 反相器的 dc 特性对这些变化相当不敏感，因此该门能在一

① 此处注意公式（5.10）对于这一具体例子并不完全成立。细心的读者将发现对于现有的工作情况 PMOS 工作在饱和模式而不是速度饱和。然而这对于结果的影响很小。

个很宽范围的工作条件下正确工作。这在图 5.7 中已很清楚，图中显示出器件尺寸的变化对反相器的开关阈值只产生很小的影响。为了对该门所具有的稳定性进一步确认，我们将典型器件替换为最坏或最好情形下的器件来重新模拟电压传输特性。图 5.11 画出了这两种极端的情形：一个比预想要好的 NMOS 与一个很差的 PMOS 组合的情形以及相反的情形。比较所得到的曲线与典型的响应表明，该门的工作没有受任何影响，这些变化主要引起开关阈值的平移。这一特性确保了该门能在一个很宽范围的条件下工作，这也是静态 CMOS 门得以普遍使用的主要原因。

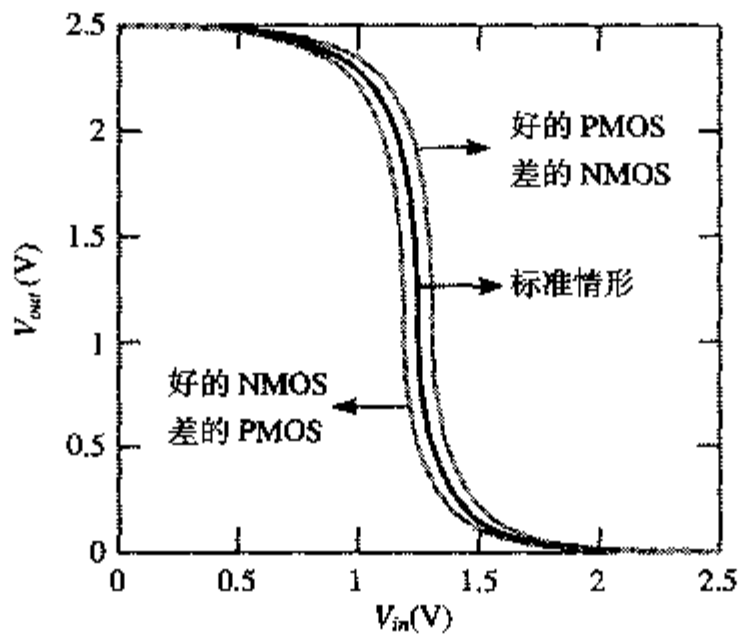


图 5.11　器件参数变化对静态 CMOS 反相器 VTC 的影响。好的器件具有较小的栅氧厚度（减小 3 nm）、较小的长度（减小 25 nm）、较大的宽度（加大 30 nm）以及较小的阈值（减小 60 mV）。对于差的晶体管情形则相反

降低电源电压

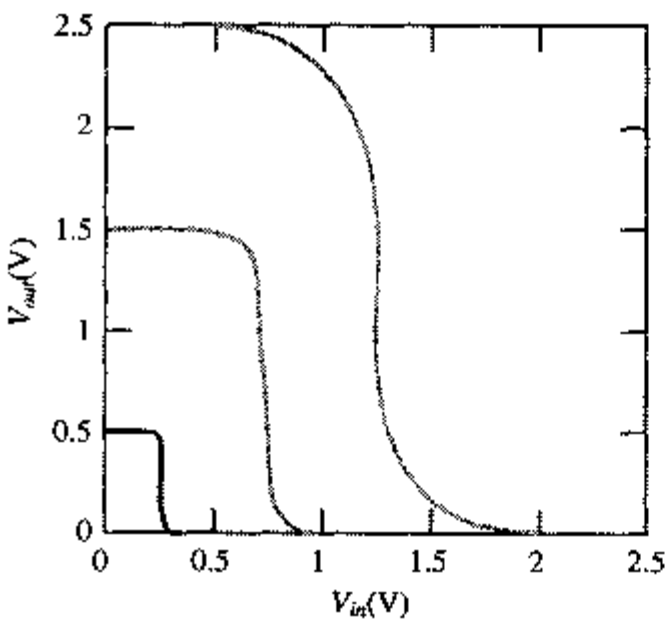
在第 3 章中我们看到工艺尺寸的连续缩小已迫使电源电压与器件尺寸按类似的比例降低。与此同时器件阈值电压实际上却保持不变。我们也许想知道这一趋势对 CMOS 反相器工作的稳定性会有什么影响。当电压降低时反相器是否仍然工作？对电源电压的降低是否存在一个可能的限制？

公式（5.10）提供了第一条线索，它告诉我们可能会发生什么，指出了反相器在过渡区的增益实际上随电源电压的降低而加大！注意，对于固定的晶体管尺寸比 r，V_M 近似地正比于 V_{DD}。画出不同电源电压时的（归一化）VTC 不仅证实了这一猜想，而且说明反相器在电源电压接近构成它的晶体管的阈值电压时仍能很好地工作［见图 5.12（a）］。当电源电压为 0.5 V 时——只比晶体管的阈值高 100 mV，过渡区的宽度只是电源电压的 10%（最大增益为 35），而电源电压为 2.5 V 时，它加大到 17%。那么既然这可以改善 dc 特性，为什么我们不使所有的数字电路都选择在这样低的电源电压下工作呢？可以想到有三个重要的理由：

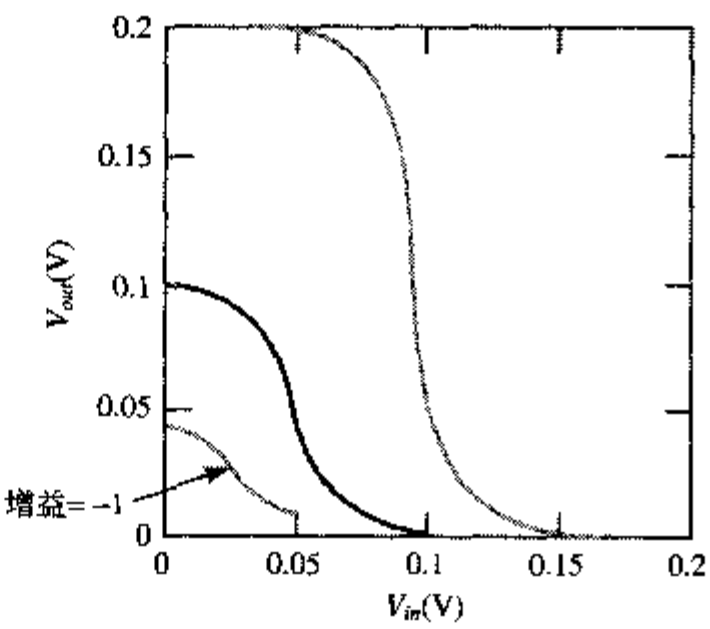
- 不加区分地降低电源电压虽然对减少能耗有正而影响，但它绝对会使门的延时加大。这将在下一节讨论。
- 一旦电源电压和本征电压（阈值电压）变得可比拟，dc 特性对器件参数（如晶体管阈值）的变化就变得越来越敏感。
- 降低电源电压意味着减小信号摆幅。虽然这通常可以帮助减少系统的内部噪声（如由串扰引起的噪声），但它也使设计对并不减小的外部噪声源更加敏感。

为了深入了解对电压降低可能存在的限制，我们在图 5.12（b）中画出了同一个反相器甚至在更低的电源电压 200 mV、100 mV 和 50 mV 时的电压传输特性。晶体管的阈值仍保持在同一值。

令人奇怪的是尽管电源电压不足以大到使晶体管导通，但我们仍然得到了一个反相器的特性！这从晶体管的亚阈值工作中可以得到解释。亚阈值电流足以使该门在低电平和高电平之间切换，并提供足够的增益从而得到可接受的VTC！开关电流值很低决定了反相器工作得非常慢，但这也许在某些应用（例如手表）中可以接受。



（a）降低 V_{DD} 改善了增益



（b）对非常低的电源电压传输特性变差

图 5.12 CMOS 反相器的 VTC 与电源电压的关系（0.25 μm CMOS 工艺）

当接近100 mV 时，我们开始看到门的特性变得很差。V_{OL} 和 V_{OH} 不再等于电源的两个电平，并且过渡区的增益接近1。为了能达到足够的增益以用于数字电路，必须使电源至少等于第3章所说的热电势，即$\phi_T = kT/q$（室温时为25 mV）的两倍［Swanson72］。因此当低于这一电压时，热噪声也会成为一个问题而可能引起不可靠的工作。我们把这一关系表示为：

$$V_{DDmin} > 2 \ldots 4\frac{kT}{q} \tag{5.11}$$

公式（5.11）代表了电源电压降低的实际限制。这意味着使CMOS 反相器工作在100 mV 以下的惟一方法是降低环境温度，或者换言之，冷却该电路。

思考题 5.3 CMOS 反相器的最小电源电压

一旦电源电压低于阈值电压，晶体管就工作在亚阈值区域，并显示出指数的电流-电压关系［如公式（3.39）所示］。推导出在这些条件下反相器增益的表达式（假设对称的 NMOS 和 PMOS 管，并且 $V_M = V_{DD}/2$ 时增益最大）。所得到的表达式表明最小的电压与晶体管的斜率系数 n 有关：

$$g = -\left(\frac{1}{n}\right)\left(e^{V_{DD}/2\phi_T} - 1\right) \tag{5.12}$$

根据这个表达式，增益在 $V_{DD} = 48$ mV 时降至−1（当 $n = 1.5$ 及 $\phi_T = 25$ mV 时）。

5.4 CMOS 反相器的性能：动态特性

前面的定性分析表明CMOS反相器的传播延时取决于它分别通过PMOS和NMOS管充电和放电负载电容 C_L 所需要的时间。这一结果说明使 C_L 尽可能小是实现高性能CMOS 电路的关键，因此在着手深入分析该门的传播延时之前有必要先研究一下负载电容的主要组成部分。除这一细节分析之外，本节还总结了设计者可以用来优化反相器性能的技术。

5.4.1　计算电容值

如果对一个 MOS 电路中的每一个电容分别进行考虑，那么对这个电路进行手工分析事实上是不可能的。这一情况因在 MOS 晶体管模型中存在许多非线性电容而更加严重。为了使分析容易进行，我们假设所有的电容一起集总成一个单个的电容 C_L，它处于 V_{out} 和 GND 之间。注意，这是实际情形相当程度的简化，甚至在简单反相器的情形中也是如此。

图 5.13 为一对串联反相器的电路图。它包括了影响节点 V_{out} 瞬态响应的所有电容。先假设输入 V_{in} 由一个上升和下降时间均为零的理想电压源所驱动。只考虑连至输出节点上的电容时，C_L 可以分解为以下几部分。

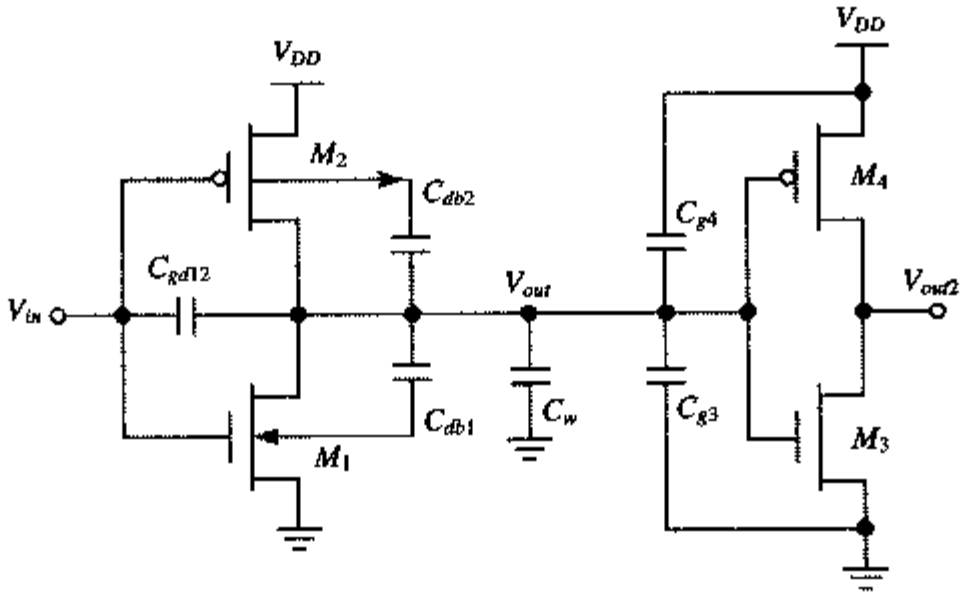


图 5.13　影响一对串联反相器动态特性的寄生电容

栅漏电容 C_{gd12}

在输出过渡的前半部(至50%的点)，M_1 和 M_2 不是断开就是处在饱和模式。在这些情况下 C_{gd12} 只包括 M_1 和 M_2 的覆盖电容。MOS 晶体管的沟道电容在这里不起作用，因为它完全处在栅和体(断开时)或栅和源(饱和时)之间(见第3章)。

集总电容模型要求用接地电容来代替浮空的栅漏电容，这是通过考虑所谓的密勒效应来实现的。在由低至高或由高至低的过渡中，栅漏电容两端的电压向相反的方向变化(见图5.14)。因此在这一浮空电容上的电压变化是实际输出电压摆幅的两倍。为了在输出节点上出现同样的负载，接地电容的值必须是浮空电容的两倍。

我们用以下公式来计算栅漏电容：$C_{gd}=2\,C_{GD0}W$(这里，C_{GD0} 是在 SPICE 模型中采用的每单位宽度的覆盖电容)。关于密勒效应的深入讨论请参见其他教科书，如[Sedra87, p.57]①。

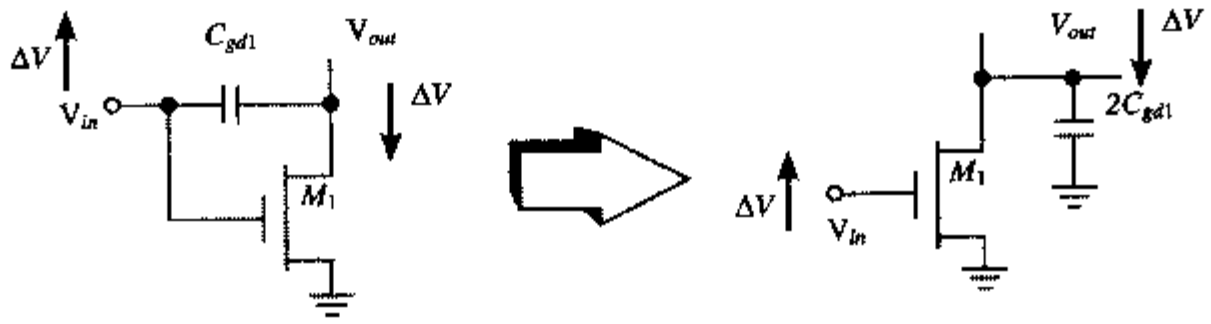


图 5.14　密勒效应——一个在其两端经历大小相同但相位相反的电压摆幅的电容可以用一个两倍于该电容值的接地电容来代替

① 这里讨论的密勒效应是一般模拟情况的简化形式。在一个数字反相器中，在输入和输出间的大信号增益总是等于−1。

扩散电容 C_{db1} 和 C_{db2}

在漏和体之间的电容来自反向偏置的*pn*结。可惜的是，这样的一个电容是高度非线性的并且在很大程度上取决于所加的电压。在第 3 章论证了进行简化分析的最好办法是用一个线性电容来代替非线性电容，使这个线性电容在所关注电压范围内的电荷变化与非线性电容的情况相同。可引入一个乘数因子 K_{eq} 来联系线性化的电容和零偏置条件下的结电容的值：

$$C_{eq} = K_{eq} C_{j0} \tag{5.13}$$

这里，C_{j0} 是在零偏置条件下单位面积的结电容。为了方便起见，我们在此重新列出方程（3.11）：

$$K_{eq} = \frac{-\phi_0^m}{(V_{high} - V_{low})(1-m)}[(\phi_0 - V_{high})^{1-m} - (\phi_0 - V_{low})^{1-m}] \tag{5.14}$$

式中，ϕ_0 是内建结电势而 m 是结的梯度系数。注意，反向偏置结的结电势定义为负值。

例 5.3　2.5 V CMOS 反相器的 K_{eq}

考虑图5.13 中用通用0.25 μm CMOS 工艺设计的反相器。与这一工艺相关的电容参数总结于表3.5 中。

让我们先来分析NMOS 管（见图5.13 中的 C_{db1}）。传播延时定义为在输入和输出翻转的50%之间的时间。对于CMOS 反相器，达到翻转的 50%的时刻是指 V_{out} 达到 1.25 V 的时刻，因为输出电压在电源的两条轨线电压之间摆动或者说等于2.5 V。因此对由高至低的翻转，我们在{2.5 V, 1.25 V }的区间上线性化结电容，而对由低至高的翻转在{0, 1.25 V }的区间上线性化结电容。

在输出由高至低的翻转期间，V_{out} 最初等于2.5 V。由于NMOS 器件的体连至*GND*，这相当于在漏结上有 2.5 V 的反向电压，即 $V_{high} = -2.5$ V。在输出达到幅值的 50%点处，$V_{out} = 1.25$ V 或者说 $V_{low} = -1.25$ V。由公式（5.14）求底板和侧壁部分的扩散电容得到下列数据：

$$\text{底板：}\quad K_{eq}\,(m = 0.5, \phi_0 = 0.9) = 0.57$$
$$\text{侧壁：}\quad K_{eqsw}\,(m = 0.44, \phi_0 = 0.9) = 0.61$$

在由低至高的翻转期间，V_{low} 和 V_{high} 分别等于0 V 和–1.25 V，这使 K_{eq} 值更高：

$$\text{底板：}\quad K_{eq}\,(m = 0.5, \phi_0 = 0.9) = 0.79$$
$$\text{侧壁：}\quad K_{eqsw}\,(m = 0.44, \phi_0 = 0.9) = 0.81$$

PMOS 管表现出相反的特性，因为它的衬底连至 2.5 V 电压。所以，对于由高至低的翻转（$V_{low} = 0$, $V_{high} = -1.25$ V），我们有：

$$\text{底板：}\quad K_{eq}\,(m = 0.48, \phi_0 = 0.9) = 0.79$$
$$\text{侧壁：}\quad K_{eqsw}\,(m = 0.32, \phi_0 = 0.9) = 0.86$$

最后，对于由低至高的翻转（$V_{low} = -1.25$ V, $V_{high} = -2.5$ V），我们有：

$$\text{底板：}\quad K_{eq}\,(m = 0.48, \phi_0 = 0.9) = 0.59$$
$$\text{侧壁：}\quad K_{eqsw}\,(m = 0.32, \phi_0 = 0.9) = 0.7$$

采用这一方法，结电容可以用一个线性电容来代替，因而可以当做和任何其他器件电容一样

来处理。线性化的结果会使电压和电流波形有微小误差。但这一简化对逻辑延时没有明显的影响。

连线电容 C_w

由连线引起的电容取决于连线的长度和宽度，并且与扇出离开驱动门的距离以及扇出门的数目有关。正如第 4 章中所论述的，这一部分电容的重要性随着工艺尺寸的缩小日益增加。

扇出的栅电容 C_{g3} 和 C_{g4}

我们假设扇出电容等于负载门 M_3 和 M_4 总的栅电容，因此，

$$\begin{aligned} C_{fan\text{-}out} &= C_{gate}(\text{NMOS}) + C_{gate}(\text{PMOS}) \\ &= (C_{GSOn} + C_{GDOn} + W_n L_n C_{ox}) + (C_{GSOp} + C_{GDOp} + W_p L_p C_{ox}) \end{aligned} \tag{5.15}$$

这一表达式在两方面简化了实际情形：

- 它假设栅电容的所有部分都连在 V_{out} 和 GND（或 V_{DD}）之间，并且忽略了栅漏电容上的密勒效应。这对精度的影响比较小，因为我们可以很有把握地假设所连接的门在达到50%点之前是不会翻转的，因而 V_{out2} 在我们所关注的时间内保持不变。
- 第二方面的近似是认为所连接的门的沟道电容在我们所关注的时间内保持不变。这与我们在第 3 章中所发现的情况不完全一样。器件的总沟道电容与器件的工作模式有关，并且在约（2/3）WLC_{ox}（饱和状态）至整个 WLC_{ox}（线性及截止状态）之间变化。总栅电容的下降也恰好发生在晶体管导通之前，如图3.31 所示。在过渡过程的前半段，可以假设其中一个负载器件一直处于线性模式，而另一个则从截止模式进入饱和状态。忽略电容的这一变化会使估计值偏保守并有大约10%的误差，但这对于一阶分析是可以接受的。

例 5.4　一个 0.25 μm CMOS 反相器的电容

用0.25 μm CMOS 工艺设计一个最小尺寸的对称CMOS 反相器。图 5.15 是它的版图。电源电压 V_{DD} 设为2.5 V。从版图中可以算出晶体管的尺寸、扩散区的面积和周长。这些数据总结在表5.1 中。作为一个例子，我们将推导 NMOS 管的漏区面积和周长。漏区是由金属-扩散层接触孔（其面积为 $4 \times 4\lambda^2$）及在接触孔与栅之间的矩形区域（其面积为 $3 \times 1\ \lambda^2$）构成的，其总面积为 $19\ \lambda^2$，即 $30\ \mu m^2$（当 $\lambda = 0.125\ \mu m$ 时）。漏区的周长比较复杂，包括以下部分（反时针方向）：$5 + 4 + 4 + 1 + 1 = 15\lambda$，即 $PD = 15 \times 0.125 = 1.875\ \mu m$。注意，漏区周长没有包括在栅一侧的部分，因为它没有被看做侧壁的一部分。PMOS 管的漏区面积和周长可以类似地推导（矩形形状使这一计算过程变得非常简单）：$AD = 5 \times 9\lambda^2 = 45\lambda^2$，即 $0.7\ \mu m^2$；$PD = 5 + 9 + 5 = 19\ \lambda$，即 $2.375\ \mu m$。

表 5.1　反相器的晶体管数据

	W/L	AD (μm^2)	PD (μm)	AS (μm^2)	PS (μm)
NMOS	0.375/0.25	0.3 ($19\lambda^2$)	1.875 (15λ)	0.3 ($19\lambda^2$)	1.875 (15λ)
PMOS	1.125/0.25	0.7 ($45\lambda^2$)	2.375 (19λ)	0.7 ($45\lambda^2$)	2.375 (19λ)

这一实际的信息可以与前面推导的近似式联合起来得到 C_L 的估计值。我们的通用工艺的电容参数总结在表3.5 中，为方便起见在此重新列出：

覆盖电容：$COD0$ (NMOS) = 0.31 fF/μm；CGDO (PMOS) = 0.27 fF/μm

底板结电容：CJ (NMOS) = 2 fF/μm^2；CJ (PMOS) = 1.9 fF/μm^2

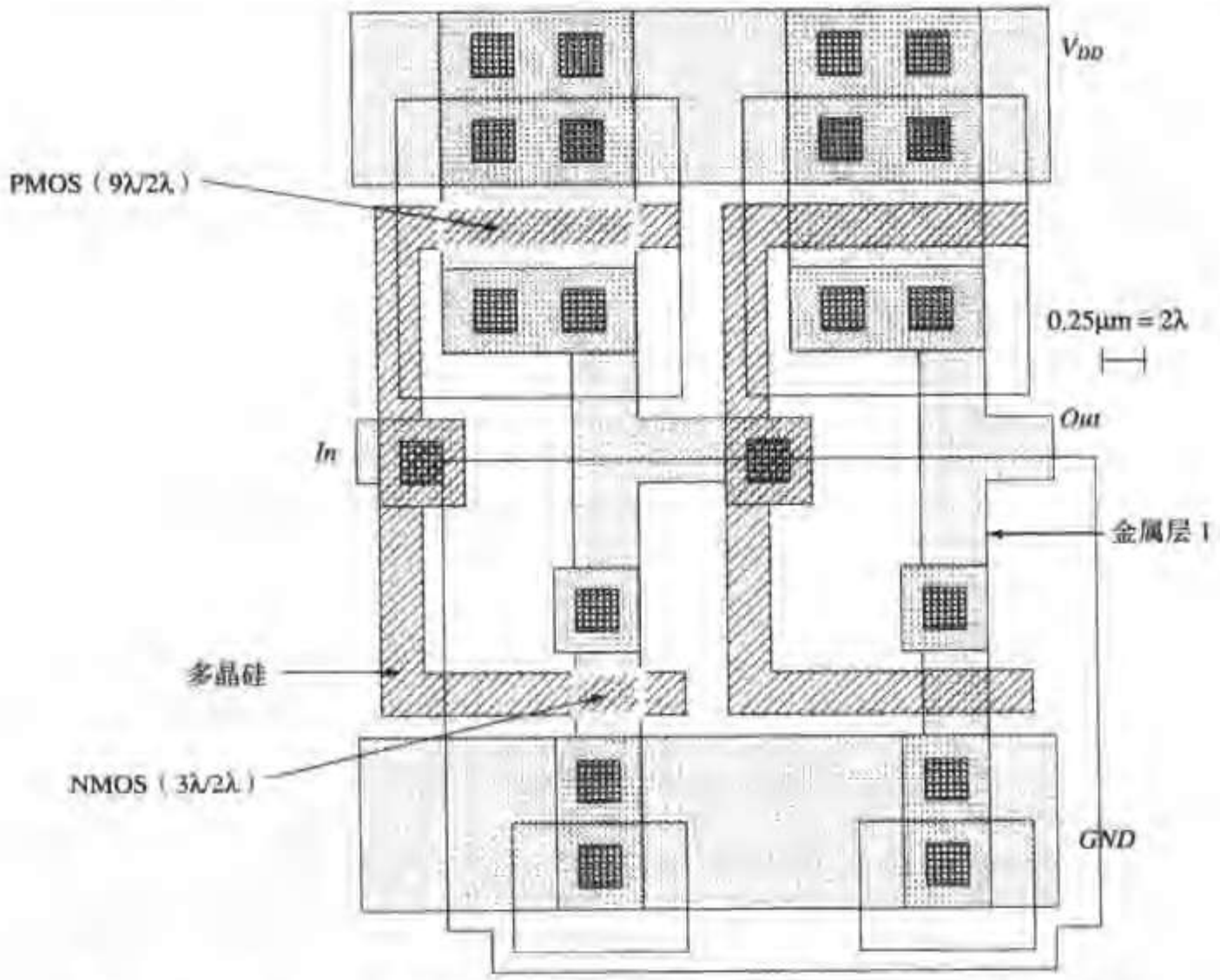


图 5.15 采用 CMOS 设计规则的两个串联的最小尺寸反相器的版图（见彩图 6）

侧壁结电容：CJSW (NMOS) = 0.28 fF/μm；CJSW (PMOS) = 0.22 fF/μm

栅电容：C_{ox} (NMOS) = C_{ox} (PMOS) = 6 fF/μm^2

最后，我们还应当考虑用金属层 1（metal 1）和多晶实现的连接门的导线所引起的电容。一个版图提取程序一般可以给出这一寄生电容的精确值。仔细阅读版图可以帮助我们进行一阶估计。由此可以得到不在有源扩散区上的 metal 1 和多晶导线的面积分别等于 42λ^2 和 72λ^2。借助表 4.2 互连参数的帮助，我们可以求出线电容（注意在这一简单计算中忽略了边缘电容；由于导线很短，它的作用与其他因素相比可以忽略）：

$$C_{wire} = 42/8^2\ \mu m^2 \times 30\ aF/\mu m^2 + 72/8^2\ \mu m^2 \times 88\ aF/\mu m^2 = 0.12\ fF$$

把所有各部分合在一起的结果总结在表 5.2 中。我们用例 5.3 推导的 K_{eq} 值来计算扩散电容。注意，负载电容几乎平均地分配在它的两个主要部分：即本征电容（由扩散电容和覆盖电容组成）和外部负载电容（由导线和所连接的门引起）。

表 5.2 C_L 的组成（由高至低和由低至高的过渡情况）

电容	表达式	值（fF）（H→L）	值（fF）（L→H）
C_{gd1}	$2\,CGD0_nW_n$	0.23	0.23
C_{gd2}	$2\,CGD0_pW_p$	0.61	0.61
C_{db1}	$K_{eqn}AD_nCJ+K_{eqnwn}PD_nCJSW$	0.66	0.90
C_{db2}	$K_{eqp}AD_pCJ+K_{eqpwp}PD_pCJSW$	1.5	1.15

（续表）

电容	表达式	值（fF）（H→L）	值（fF）（L→H）
C_{g3}	$(CGD0_n+CGSO_n)W_n+C_{ox}W_nL_n$	0.76	0.76
C_{g4}	$(CGD0_p+CGSO_p)W_p+C_{ox}W_pL_p$	2.28	2.28
C_w	提取参数	0.12	0.12
C_L	Σ	6.1	6.0

5.4.2　传播延时：一阶分析

计算反相器传播延时的一种方法是对电容的充（放）电电流积分。由此得到表达式：

$$t_p = \int_{v_1}^{v_2} \frac{C_L(v)}{i(v)} dv \tag{5.16}$$

这里，i 是充（放）电电流，v 是电容上的电压，而 v_1 和 v_2 分别是初始和最终电压。确切求解这一方程是很困难的，因为 $C_L(v)$ 和 $i(v)$ 都是 v 的非线性函数。因此我们再回到图 5.6 中介绍的简化的反相器开关模型，以推导适合于手工分析传播延时的合理近似公式。对导通电阻取决于电压的关系及负载电容的考虑是通过分别用一个常数线性元件代替它们来处理的，常数线性元件的值取它在所关注时间间隔内的平均值。前一节已精确推导了负载电容的这个值。MOS 晶体管的平均导通电阻的表达式已在例3.8 中推导过，这里为方便起见再次列出：

$$R_{eq} = \frac{1}{V_{DD}/2} \int_{V_{DD/2}}^{V_{DD}} \frac{V}{I_{DSAT}(1+\lambda V)} dV \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}\lambda V_{DD}\right) \tag{5.17}$$

其中

$$I_{DSAT} = k'\frac{W}{L}\left((V_{DD}-V_T)V_{DSAT}-\frac{V_{DSAT}^2}{2}\right)$$

现在推导所得电路的传播延时就很容易，它只不过是分析一阶线性 RC 电路，完全与例4.5 的过程相同。从例 4.5 中我们得知，由一个电压阶跃激励时，这样一个电路的传播延时正比于由这个电路的下拉电阻和负载电容形成的时间常数。因此，

$$t_{pHL} = \ln(2)R_{eqn}C_L = 0.69R_{eqn}C_L \tag{5.18}$$

同样，可以得到由低至高翻转的传播延时。我们可以写出：

$$t_{pLH} = 0.69R_{eqp}C_L \tag{5.19}$$

式中，R_{eqp} 是PMOS 管在所关注时间内的等效导通电阻。这一分析假设等效的负载电容对于由高至低及由低至高的翻转是完全相同的。在前一节的例子中已说明这是近似的情形。反相器的总传播延时定义为这两个值的平均：

$$t_p = \frac{t_{pHL}+t_{pLH}}{2} = 0.69C_L\left(\frac{R_{eqn}+R_{eqp}}{2}\right) \tag{5.20}$$

人们常常希望一个门对于上升和下降输入具有相同的传播延时。这一状况可以通过使 NMOS 和

PMOS 晶体管的导通电阻近似相等来实现。记住，这一条件与对称VTC所要求的条件是一样的。

例 5.5 一个 0.25 μm CMOS 反相器的传播延时

我们用公式（5.18）和公式（5.19）来推导图5.15中CMOS反相器的传播延时。在例5.4中已计算了负载电容C_L，而通用0.25 μm CMOS工艺的晶体管的等效导通电阻可由表3.3得到。对于2.5 V的电源电压，典型的NMOS和PMOS晶体管的导通电阻分别等于13 kΩ和31 kΩ。由版图我们确定NMOS晶体管的W与L之比为1.5，而 PMOS 为4.5。假设版图所画尺寸与实际有效尺寸的差别很小可以忽略，这就得到了以下的延时值：

$$t_{pHL} = 0.69 \times \left(\frac{13\ \text{k}\Omega}{1.5}\right) \times 6.1\ \text{fF} = 36\ \text{ps}$$

$$t_{pLH} = 0.69 \times \left(\frac{31\ \text{k}\Omega}{4.5}\right) \times 6.0\ \text{fF} = 29\ \text{ps}$$

及

$$t_p = \left(\frac{36 + 29}{2}\right) = 32.5\ \text{ps}$$

这一分析的精确性可以通过对由图5.15 的版图提取的电路图进行 SPICE 瞬态模拟来检验。计算得到的瞬态响应画在图5.16中，它确定了传播延时对HL（由高至低）翻转和LH（由低至高）翻转分别为39.9 ps 和31.7 ps。如果考虑到手工分析推导过程中有许多简化，那么它的分析结果还是很好的。特别要注意在模拟输出信号中的过冲。这些是由反相器晶体管的栅漏电容造成的，它们在晶体管对输入变化开始响应之前就直接把输入节点上变化陡峭的阶跃电压耦合到输出上。这些过冲显然对门的性能有负面的影响，同时也解释了为什么模拟的延时要比估计的大。

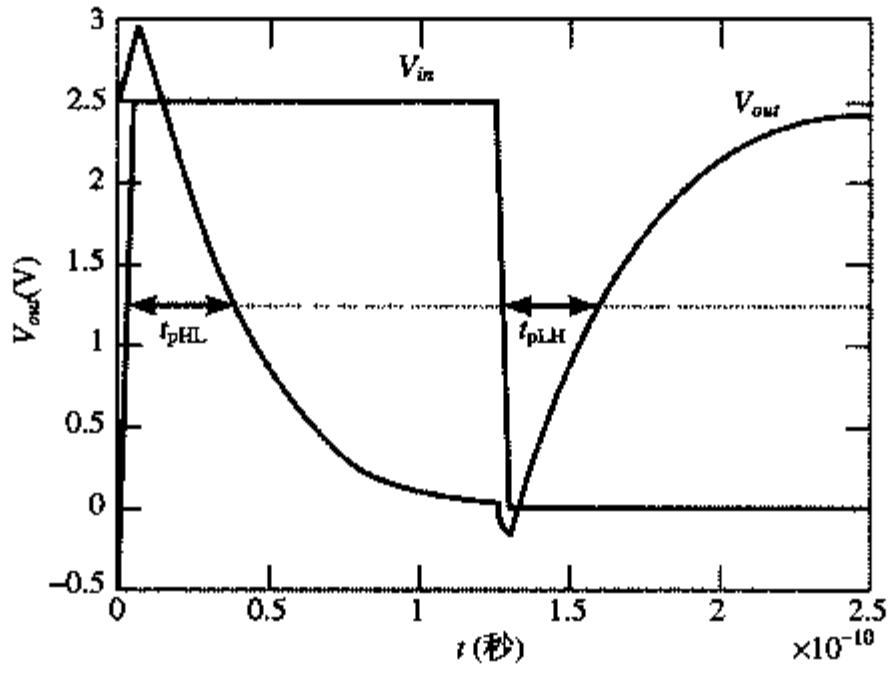


图 5.16 模拟得到的图 5.15 反相器的瞬态响应

警告：这一例子也许给出这样的结论：手工分析总是能很接近实际的响应。但并不一定是这样的。在一阶和较高阶的模型之间常常可以看到存在很大的偏差。手工分析的目的是要得到对电路特性的基本了解并确定主要的参数。在需要定量数据的时候，一个细节的模拟是不可缺少的。只能把上面的例子当做是碰巧而已。

至此，一个设计者明显要问的问题是如何处理或优化门的延时。为了回答这个问题，有必要

展开延时公式中的R_{eq}以显示出决定延时的参数。将公式（5.18）和公式（5.17）联合起来，并暂且假设忽略沟长调制系数λ，于是就得到了以下t_{pHL}的表达式（同样的分析适合于t_{pLH}）：

$$t_{pHL} = 0.69\frac{3}{4}\frac{C_L V_{DD}}{I_{DSATn}} = 0.52\frac{C_L V_{DD}}{(W/L)_n k'_n V_{DSATn}(V_{DD} - V_{Tn} - V_{DSATn}/2)} \qquad (5.21)$$

在大多数设计中，电源电压都选择得足够高，所以$V_{DD} >> V_{Tn} + V_{DSATn}/2$。在这些条件下，延时实际上与电源电压无关：

$$t_{pHL} \approx 0.52\frac{C_L}{(W/L)_n k'_n V_{DSATn}} \qquad (5.22)$$

注意，这只是一阶近似，由于沟长调制系数非零，提高电源电压将使性能得到尽管很小但可以观察到的改善。这一分析在图5.17中得到了证实，图中画出了反相器的传播延时与电源电压的关系。这一结果并不奇怪，因为这条曲线事实上与图3.28曲线的形状完全一样，后者画出了MOS晶体管的等效导通电阻与V_{DD}的关系。虽然对于较高的V_{DD}值，延时对于电源电压的变化较不敏感，但当V_{DD}接近$2V_T$时将看到延时开始迅速增加。因此如果达到高性能是主要的设计目标，那么显然应当避免在这个工作区工作。

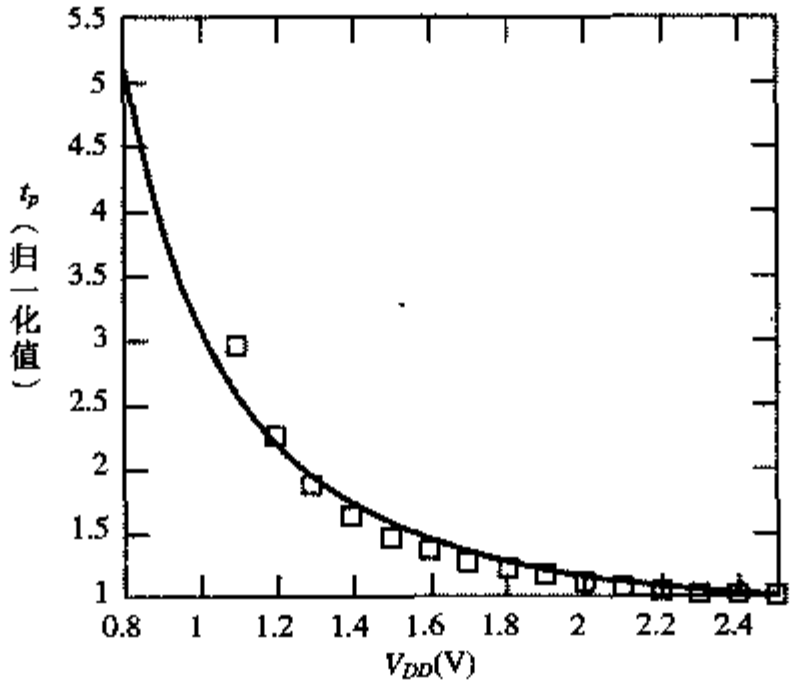


图 5.17　CMOS 反相器传播延时与电源电压的关系（归一至2.5 V时的延时）。小方块表示由公式（5.21）预见的延时值。注意这一公式只在器件速度饱和时成立。因此在低电源电压时会有偏差

■

设计技术

由以上讨论我们得知可以用以下方式减小一个门的传播延时：

- 减小C_L：注意这个负载电容由三个主要部分组成：门本身的内部扩散电容、互连线电容和扇出电容。细致的版图设计有助于减少扩散电容和互连线电容。优秀的设计实践要求漏扩散区的面积越小越好。
- 增加晶体管的W/L比：这是设计者手中最有力和最有效的性能优化工具。但是在采用这一

方法时要当心。增加晶体管尺寸也增加扩散电容，因而增加了C_L。事实上，一旦本征电容（即扩散电容）开始超过由连线和扇出形成的外部负载，增加门的尺寸就不再对减少延时有帮助。它只是加大了门的面积，这称为自载效应。此外，较宽的晶体管具有较大的栅电容，这就增加了驱动门的扇出系数，从而又反过来影响它的速度。

- 提高V_{DD}：如图5.17所示，一个门的延时可以通过改变电源电压来调整。这一灵活性使设计者可以用能量损耗来换取性能，正如我们在后面一节中将要看到的那样。然而，增加电源电压超过一定程度后改善就会非常有限，因而应当避免。同时对可靠性方面的考虑（氧化层的击穿，热电子效应）也迫使在深亚微米工艺中对电源电压要规定严格的上限。

思考题 5.4 传播延时与充（放）电电流的关系

至此，我们已把传播延时表示成晶体管等效电阻的函数。另一种方法是用一个与在所关注的时间间隔内平均充（放）电电流值相等的电流源来代替该晶体管。试运用这一不同的方法来推导传播延时的表达式。

5.4.3 从设计角度考虑传播延时

从前面推导出的延时表达式中可以得出一些有意义的设计综合考虑原则。最重要的是，它们可以形成确定晶体管尺寸的一般方法，而这一方法将证明是极为有用的。

NMOS 与 PMOS 的比

至今我们一直使PMOS管较宽，以使它的电阻与下拉的NMOS管匹配。这通常要求PMOS和NMOS的宽度比在3～3.5之间。采用这一方法的目的是设计一个具有对称VTC的反相器并使由高至低与由低至高的传播延时相等。然而这并不意味着这一比值也得到最小的总传播延时。如果对称性和噪声容限不是主要的考虑因素，那么实际上有可能通过减小PMOS器件的宽度来加快反相器的速度！

以上说法的理由是，虽然使PMOS较宽因充电电流的增加而改善了反相器的t_{pLH}，但它也由于产生较大的寄生电容而使t_{pHL}变差。当这两个相反的效应都存在时，必定存在一个晶体管的宽度比使反相器的传播延时最优（最小）。

这一优化的比值可以用简单的解析方法来推导。考虑两个完全相同的CMOS反相器相串联。第一个门的负载电容可近似为：

$$C_L = (C_{dp1} + C_{dn1}) + (C_{gp2} + C_{gn2}) + C_W \qquad (5.23)$$

式中，C_{dp1}和C_{dn1}是第一个反相器PMOS和NMOS晶体管的漏扩散电容，而C_{gp2}和C_{gn2}为第二个反相器的栅电容，C_W代表连线电容。

当PMOS器件为NMOS器件的β倍时[$\beta = (W/L)_p/(W/L)_n$]，所有的晶体管电容也以近似相同的比例加大，即$C_{dp1} \approx \beta C_{dn1}$，及$C_{gp2} \approx \beta C_{gn2}$。于是公式（5.23）可以重写成：

$$C_L = (1 + \beta)(C_{dn1} + C_{gn2}) + C_W \qquad (5.24)$$

基于公式（5.20）可以推导出以下传播延时的表达式：

$$t_p = \frac{0.69}{2}((1+\beta)(C_{dn1}+C_{gn2})+C_W)\left(R_{eqn}+\frac{R_{eqp}}{\beta}\right)$$
$$= 0.345((1+\beta)(C_{dn1}+C_{gn2})+C_W)R_{eqn}\left(1+\frac{r}{\beta}\right) \quad (5.25)$$

这里，$r\ (=R_{eqp}/R_{eqn})$代表尺寸完全相同的PMOS和NMOS晶体管的电阻比。令$\frac{\partial t_p}{\partial \beta}$等于零可以求出β的最大值，即：

$$\beta_{opt} = \sqrt{r\left(1+\frac{C_W}{C_{dn1}+C_{gn2}}\right)} \quad (5.26)$$

这意味着当导线电容可以忽略时（即$C_{dn1}+C_{gn2} \gg C_W$），β_{opt}等于$\sqrt{r}$，这不同于在非串联情形时通常采用的比值r。如果导线电容占主导地位，那么应当采用较大的β值。这一分析令人奇怪的结果是当以对称性及噪声容限为代价时，较小的器件尺寸（因而较小的设计面积）得到了较快的设计。

例 5.6　确定以相同门为负载的 CMOS 反相器的尺寸

再次考虑我们标准的设计例子。由等效电阻值（见表3.3）发现比值β为2.4 (= 31 kΩ/13 kΩ)时将得到对称的瞬态响应。而现在由公式（5.26）可预见到最优性能的器件比值应当等于 1.6。这些结果可从图5.18得到验证，它画出了模拟得到的传播延时与晶体管比值β的关系。该图清楚地显示了改变β将如何改变t_{pLH}和t_{pHL}值的相对大小。最优点发生在β = 1.9附近，这比预见的要高些。同时可以预见到在β = 2.4处上升和下降延时相同。当最坏情形下的延时是主要考虑因素时，这是一个所希望的工作点[①]。

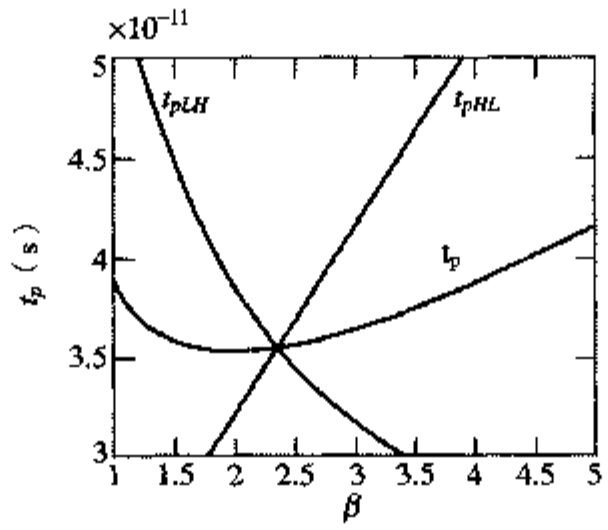


图 5.18　CMOS 反相器的传播延时与 PMOS 对 NMOS 管比值β的关系

考虑性能时反相器尺寸的确定

在这个分析中，我们假设一个对称反相器，即它的PMOS和NMOS尺寸使上升和下降延时相同。这一反相器的负载电容可以分为本征和外部两个部分，即$C_L = C_{int} + C_{ext}$。C_{int}代表反相器的自载即本征输出电容，它与NMOS和PMOS管的扩散电容以及栅漏覆盖（密勒）电容有关。C_{ext}是外部负载电容，它来自扇出和导线电容。假设R_{eq}代表门的等效电阻，我们可以把传播延时表示为：

① 你也许会奇怪为什么我们一直不把上升和下降延时中最差的一个作为衡量一个门的主要性能指标。当把反相门串联起来构成一个较为复杂的逻辑电路时，很快就会发现，把这两个延时平均起来是一个较为有意义的衡量标准。因为一个门的上升过渡之后紧接着便是下一个门的下降过渡。

$$\begin{aligned} t_p &= 0.69R_{eq}(C_{int}+C_{ext}) \\ &= 0.69R_{eq}C_{int}(1+C_{ext}/C_{int}) = t_{p0}(1+C_{ext}/C_{int}) \end{aligned} \tag{5.27}$$

式中，$t_{p0}=0.69R_{eq}C_{int}$代表反相器的负载只是其本征电容（$C_{ext}=0$）时的延时，称为本征延时或无负载延时。

下一个问题是晶体管的尺寸是如何影响门的性能的。为了回答这个问题，我们必须建立起公式（5.27）中的各种参数和尺寸系数S之间的关系。尺寸系数S把反相器的晶体管尺寸与一个参考门（通常是一个最小尺寸的反相器）的晶体管尺寸联系起来。本征电容C_{int}包括扩散电容及密勒电容，它们都正比于晶体管的宽度。因此，$C_{int}=SC_{iref}$。门的电阻与参考门的关系为$R_{eq}=R_{ref}/S$。我们现在可以把公式（5.27）重写成：

$$\begin{aligned} t_p &= 0.69(R_{ref}/S)(SC_{iref})(1+C_{ext}/(SC_{iref})) \\ &= 0.69R_{ref}C_{iref}\left(1+\frac{C_{ext}}{SC_{iref}}\right) = t_{p0}\left(1+\frac{C_{ext}}{SC_{iref}}\right) \end{aligned} \tag{5.28}$$

由这一分析可以得出两个重要的结论：

- 反相器的本征延时t_{p0}与门的尺寸无关，而只取决于工艺及反相器的版图。当不存在任何（外部）负载时，门的驱动强度的提高完全为随之而增加的电容所抵消。
- 使S无穷大将达到最大可能的性能改善，因为这消除了任何外部负载的影响，使延时减小到只有本征延时值。然而任何比（C_{ext}/C_{int}）足够大的尺寸系数S都会显著增加硅面积而得到类似的结果。

例 5.7 考虑性能时的器件尺寸确定

让我们考察一下例 5.5 中确定的器件尺寸所能达到的性能改善。由表 5.2 得到$C_{int}/C_{ext}\approx 1.05$（$C_{int}=3.0$fF，$C_{ext}=3.15$fF）。由此可以预见最大的性能改善为2.05。尺寸放大系数为 10 时得到的性能与这一最优性能的差距在 10%以内，再加大器件尺寸只能得到可以忽略不计的性能改善。

这一点已经为模拟结果所证实，它预见到可能达到的最大性能改善为 1.9（$t_{p0}=19.3$ ps）。从图 5.19 中我们看到$S=5$时已得到了大部分的改善，而尺寸系数大于 10 时几乎得不到任何额外的收益。

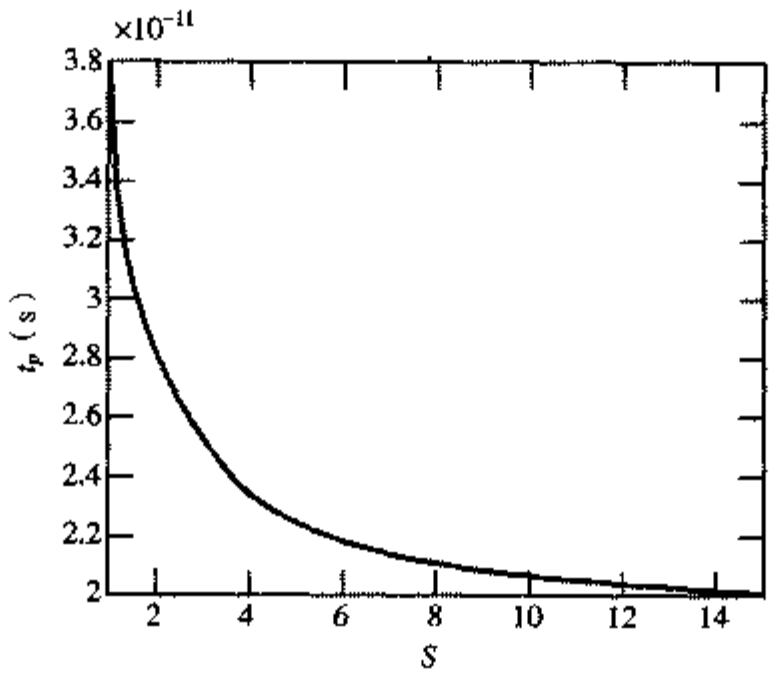


图 5.19 对固定的扇出，以相同的系数S放大 NMOS 和 PMOS 管的尺寸来提高反相器的性能（见图5.15的反相器）

确定反相器链的尺寸

虽然加大反相器的尺寸可以减小它的延时，但这也加大了它的输入电容。如孤立地确定门的尺寸而不考虑它对前级门延时的影响，则纯粹是一种脱离实际的研究。由此一个比较相关的问题是当一个门处在实际环境中时如何确定它的最优尺寸。一个简单的反相器链则是最好的研究起点。为了决定输入的负载效应，必须建立起反相器的输入栅电容C_g与本征输出电容之间的关系。这两个电容均正比于门的尺寸。因此，下列关系成立而与门的尺寸无关：

$$C_{int} = \gamma C_g \tag{5.29}$$

在公式（5.29）中，γ 是比例系数，它只与工艺有关，并且对于大多数的亚微米工艺 γ 接近于 1，这正如前面的例子所示。重新写出公式（5.28），我们得到：

$$t_p = t_{p0}\left(1 + \frac{C_{ext}}{\gamma C_g}\right) = t_{p0}(1 + f/\gamma) \tag{5.30}$$

上式表明，反相器的延时只取决于它的外部负载电容与输入电容间的比值。这一比值称为等效扇出 f。

让我们考虑图5.20的电路。我们的目的是要使通过反相器链的延时最小，其中第一个反相器（通常为最小尺寸的门）的输入电容为C_{g1}，而反相器链末端为一个固定的负载电容C_L。

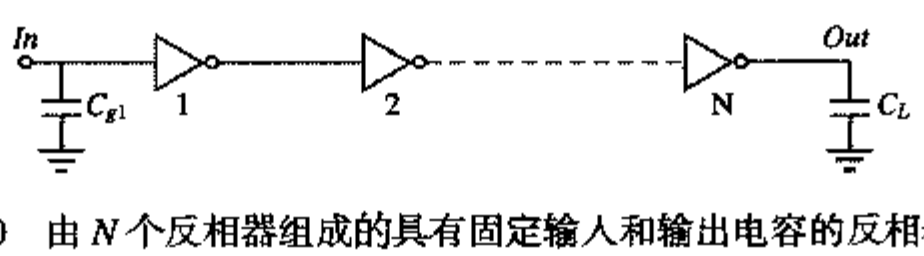


图 5.20　由 N 个反相器组成的具有固定输入和输出电容的反相器链

由第 j 级反相器的延时表达式[①]：

$$t_{p,j} = t_{p0}\left(1 + \frac{C_{g,j+1}}{\gamma C_{g,j}}\right) = t_{p0}(1 + f_j/\gamma) \tag{5.31}$$

可以推导出反相器链的总延时：

$$t_p = \sum_{j=1}^{N} t_{p,j} = t_{p0}\sum_{j=1}^{N}\left(1 + \frac{C_{g,j+1}}{\gamma C_{g,j}}\right), \text{ 其中 } C_{g,N+1} = C_L \tag{5.32}$$

这个方程含有 $N-1$ 个未知数，即 $C_{g,2}$，$C_{g,3}$…，$C_{g,N}$。通过求 $N-1$ 次偏微分并令它们都等于0，即 $\partial t_p/\partial C_{g,j}=0$，可以求得最小延时。由此得到了一组约束条件：

$$C_{g,j+1}/C_{g,j} = C_{g,j}/C_{g,j-1} \qquad \text{其中 } (j = 2 \ldots N) \tag{5.33}$$

换言之，每一个反相器的最优尺寸是与它相邻的前后两个反相器尺寸的几何平均数：

$$C_{g,j} = \sqrt{C_{g,j-1}C_{g,j+1}} \tag{5.34}$$

这意味着每个反相器的尺寸都相对于它前面反相器的尺寸放大相同的倍数 f，即每个反相器都具有相同的等效扇出（$f_j=f$），因而也就具有相同的延时。当$C_{g,1}$和 C_L 给定时，我们可以推导出

① 这一表达式忽略了导线电容，这暂时是一个合理的假设。

尺寸系数为：

$$f = \sqrt[N]{C_L/C_{g,1}} = \sqrt[N]{F} \tag{5.35}$$

以及通过该反相器链的最小延时：

$$t_p = Nt_{p0}(1 + \sqrt[N]{F}/\gamma) \tag{5.36}$$

式中，F代表该电路的总等效扇出，它等于$C_L/C_{g,1}$。注意t_p和F之间的关系是如何与反相器链的级数密切相关的。正如可以预见到的，当只存在一级时，这是一个线性关系。加入第二级将使它变为平方根关系，依次类推。现在明显的问题是对于给定的F值如何选定级数使延时最短。

选择一个反相器链的正确级数

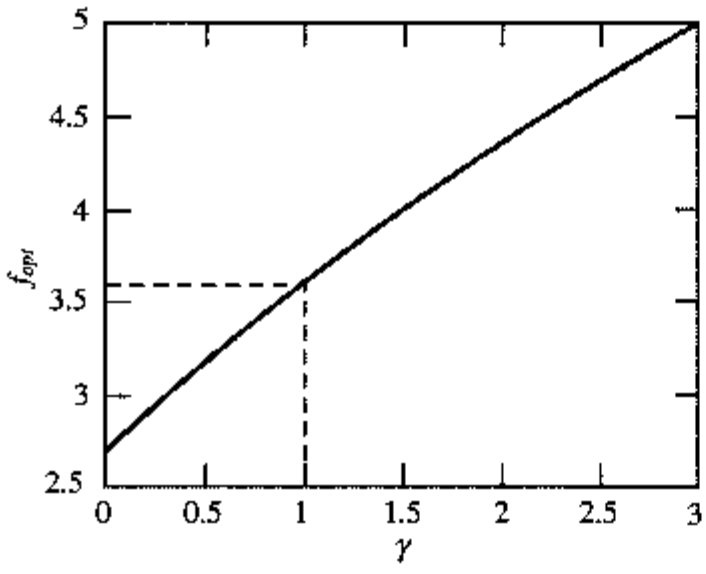
从对公式（5.36）的求值中可以看出对于给定的$F(=f^N)$在选择级数时需要综合考虑。当级数太大时，公式的第一部分（它代表了反相器级的本征延时）将占主导地位。而如果级数太小，则每一级的等效扇出变大，使公式的第二部分占主导地位。通过求最小延时表达式对级数的导数并令它为0，可以求得最优值。我们得到：

$$\gamma + \sqrt[N]{F} - \frac{\sqrt[N]{F}\ln F}{N} = 0$$

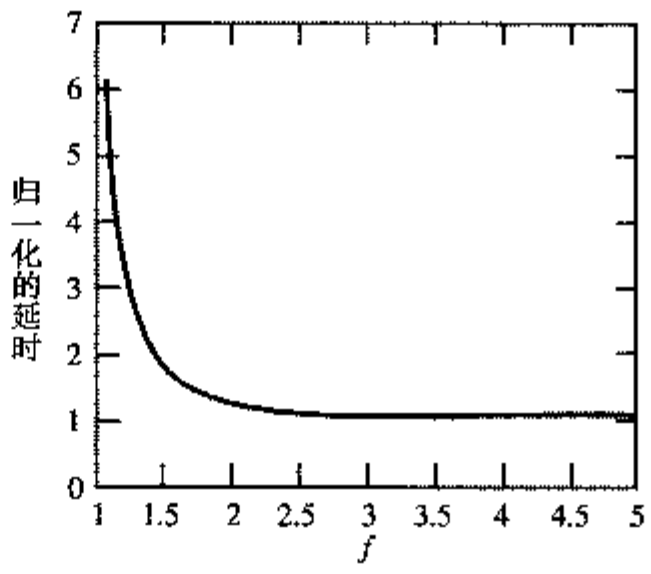
这相当于：

$$f = e^{(1+\gamma/f)} \tag{5.37}$$

方程（5.35）只有一个收敛解，即$\gamma = 0$时的解——此时忽略自载，因此负载电容只由扇出构成。在这些简化条件下可以得到最优的级数为$N = \ln(F)$，且每一级的等效扇出为$f = e = 2.718\,28$。这一优化的缓冲器设计以一种指数形式逐级放大各级尺寸，并且因此称为指数锥形［Mead80］。当包括自载时，方程（5.37）只能求数值解，其结果画在图5.21（a）中。对于$\gamma \approx 1$的典型情形，最优的锥形系数将接近于3.6。图5.21（b）画出了$\gamma = 1$时反相器链（归一化）的传播延时与等效扇出的关系。选择扇出值大于最优值并不会过多地影响延时，但能减少所要求的缓冲器级数和实现面积。一个通常的做法是选择最优的扇出为4。反之，采用过多的级数（即$f < f_{opt}$）对延时会有明显的负面影响，因而应当避免。



（a）最优的等效扇出f（或反相器的逐级放大系数）与反相器链中自载系数γ的关系



（b）自载系数γ=1时归一化的传播延时（$t_p/(t_{popt})$）与等效扇出系数的关系

图 5.21 优化反相器链中的级数

例 5.8　引入缓冲器级的影响

表5.3列出了无缓冲器的设计、两级反相器以及优化的反相器链对于不同的 F 值所对应的 $t_{p,opt}/t_{p0}$ 值（$\gamma=1$）。我们注意到在驱动非常大的电容负载时，采用串联的反相器可以达到非常明显的加速。

表 5.3　不同驱动器结构的 t_{opt}/t_{p0} 与 F 的关系

F	无缓冲器	两级反相器	反相器链
10	11	8.3	8.3
100	101	22	16.5
1000	1001	65	24.8
10 000	10 001	202	33.1

以上分析可以延伸为不仅包括反相器链，而且也包括含实际扇出的反相器网络，一个这样的例子显示在图5.22中。我们只需要调整 C_{ext} 的表达式以考虑附加的扇出系数。

思考题 5.5　确定反相器网络的尺寸

确定图5.22电路中反相器的尺寸，使在节点 *Out* 和 *In* 之间的延时最小。可以假设 $C_L=64\,C_{g,1}$。

提示：首先决定使延时最小的各个器件之间的比。应当发现以下的关系必定成立：

$$\frac{4C_{g,2}}{C_{g,1}}=\frac{4C_{g,3}}{C_{g,2}}=\frac{C_L}{C_{g,3}}$$

求门的确切尺寸（$C_{g,3}=2.52\,C_{g,2}=6.35\,C_{g,1}$）是比较容易的（注意 $2.52=16^{1/3}$）。如果直接确定反相器链的尺寸而不考虑额外的扇出，将得到尺寸系数为4而不是2.52。

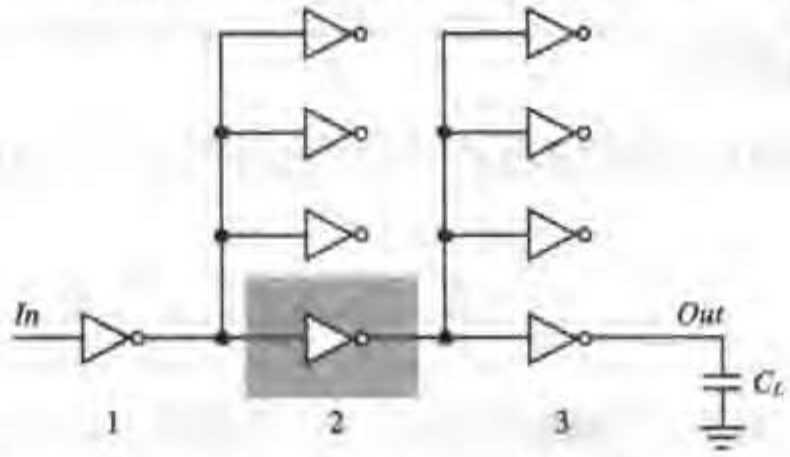


图 5.22　反相器网络。其中每个门的扇出都为 4 个门，把一个输入以树结构的形式分配给 16 个输出信号

输入信号的上升-下降时间

以上所有表达式的推导都假设了反相器的输入信号是突然从 0 变到 V_{DD} 或相反，并且假定两个晶体管中只有一个在充（放）电过程中是导通的。实际上，输入信号是逐渐变化的，而且PMOS和NMOS管会暂时同时导通一段时间。这会影响所得到的充（放）电总电流，从而影响传播延时。图5.23是通过 SPICE 模拟得到的一个最小尺寸反相器的传播延时与输入信号斜率的关系。可以看到，一旦 $t_s>t_p$，t_p 随输入斜率[①]的增加而（近似地）线性增加。

虽然可以推导出一个解析表达式来描述输入信号斜率与传播延时之间的关系，但结果会比较

① 这里的输入斜率是指 t_t，即输入的上升/下降时间。——译者注

复杂。从设计角度出发，把较慢的斜率对性能的影响直接与造成它的原因（即前面一级门的有限驱动能力）联系起来更有意义。如果后者无限大，则它输出的上升/下降时间就会是零，因此所考察的门的性能不受影响。这一步骤的优势在于它明白一个门永远不会是孤立设计的，它的性能要受扇出以及驱动其输入端的门的驱动强度的影响。这样就得到了在一个反相器链中反相器 i 传播延时的修正表达式［Hedenstierna87］：

$$t_p^i = t_{step}^i + \eta t_{step}^{i-1} \tag{5.38}$$

公式（5.38）表明，反相器 i 的传播延时等于同样的门在阶跃输入时（即输入斜率为无穷大）的延时（tistep）加上它前面一级门（i-1）的阶跃输入延时的一部分。式中比例因子 η 是一个经验常数，它的典型值约为 0.25。这个表达式的优点是它非常简单而又展示了计算复杂电路延时所需要的全部关系。

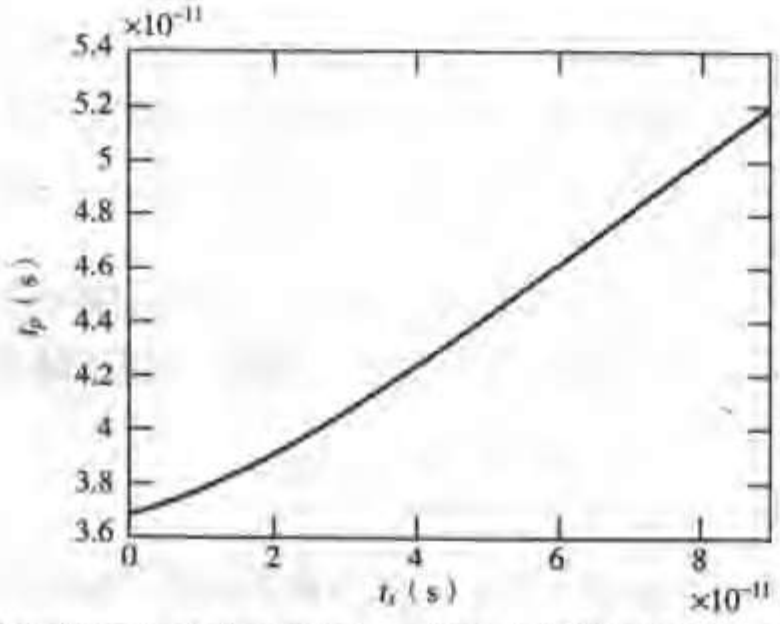


图 5.23 对于扇出为单个门的最小尺寸反相器，t_p 与输入信号斜率（10%~90%上升或下降时间）的关系

例 5.9 网络内部反相器的延时

例如，考虑图 5.22 的电路。借助于公式（5.31）和公式（5.38）可以推导出第 2 级反相器（用灰色方块标记）的延时：

$$t_{p,2} = t_{p0}\left(1 + \frac{4C_{g,3}}{\gamma C_{g,2}}\right) + \eta t_{p0}\left(1 + \frac{4C_{g,2}}{\gamma C_{g,1}}\right)$$

采用思考题 5.5 的方式对整个传播延时进行分析，得到以下最小延时所要求的尺寸的修正值：

$$\frac{4(1+\eta)C_{g,2}}{C_{g,1}} = \frac{4(1+\eta)C_{g,3}}{C_{g,2}} = \frac{C_L}{C_{g,3}}$$

如果我们假设 $\eta = 0.25$，则 f_2 和 f_1 估计为 2.47。

设计挑战

保持门的输入信号的上升时间小于或等于门的传播延时是很有利的。正如在后面将要讨论的，这不仅有利于提高性能也有利于降低功耗。使信号的上升和下降时间较小并且具有接近相等的值是高性能设计面临的主要挑战之一，这通常称为斜率工程设计。 ■

思考题 5.6 输入斜率的影响

确定降低电源电压使输入信号斜率对传播延时的影响是增加还是减少。为什么？

存在（长）互连线时的延时

至此，互连线在我们的分析中一直起着微不足道的作用。当门之间的距离进一步加大时，导线的电容和电阻就不能再被忽略，它们甚至可能主导瞬态响应。可以利用前面一章所介绍的导线模拟方法来修正前面的延时表达式以包括这些额外的影响。例4.9中的详细分析可以直接用来解决手头的问题。考虑图5.24的电路，图中一个反相器通过一条长度为L的导线驱动单个扇出。这一驱动器用单个电阻R_{dr}来代表，其大小等于R_{eqn}和R_{eqp}的平均值。C_{int}和C_{fan}分别为驱动器的本征电容和扇出门的输入电容。

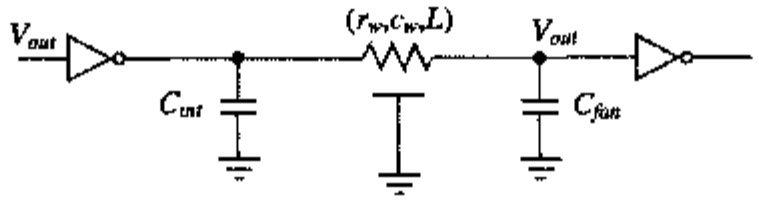


图 5.24　通过长度为 L 的导线驱动单个门的反相器

运用Elmore延时表达式可以得到电路的传播延时为：

$$\begin{aligned} t_p &= 0.69R_{dr}C_{int} + (0.69R_{dr} + 0.38R_w)C_w + 0.69(R_{dr} + R_w)C_{fan} \\ &= 0.69R_{dr}(C_{int} + C_{fan}) + 0.69(R_{dr}c_w + r_wC_{fan})L + 0.38r_wc_wL^2 \end{aligned} \qquad (5.39)$$

因子0.38说明导线事实上表现为分布延时。C_w和R_w分别代表导线的总电容和总电阻。这一延时表达式含有一个与导线长度成线性关系的部分及一个成平方关系的部分。正是后一部分使导线延时在较长导线的总延时中迅速占据支配地位。

例 5.10　有互连线时的反相器延时

考虑图5.24的电路，并假设具有例5.5的器件参数：C_{int}=3 fF，C_{fan}=3 fF，R_{dr}=0.5 (13/1.5 + 31/4.5) = 7.8 kΩ。导线由Metal1实现，宽度为0.4 μm，即所允许的最小尺寸。这得到了以下参数：c_w=92 aF/μm 和 r_w=0.19 Ω/μm（见例 4.4）。利用公式（5.39）可以计算导线长度为何值时互连延时等于单纯由器件寄生参数引起的本征延时。解下列二次方程会得到一个（有意义的）解：

$$6.6 \times 10^{-18}L^2 + 0.5 \times 10^{-12}L = 32.29 \times 10^{-12}$$

$$\text{即 } L = 65\ \mu\text{m}$$

注意，增加的延时仅来自方程的线性部分，更具体地说，是由于导线引起的额外电容。二次部分（导线分布延时）只是在导线长得多（> 7 cm）时才起主导作用。这是由于（最小尺寸）驱动晶体管的高电阻所致。如果使用较宽的管子，这两部分的对比情况就会不同。例如分析一下同一个问题，但驱动管要宽100倍。

5.5　功耗、能量和能量延时

到目前为止，我们已看到具有理想 VTC（即对称形状、全幅逻辑摆幅及高噪声容限）的静态CMOS反相器表现出极佳的稳定性，这大大简化了设计过程并打开了通往设计自动化的大门。静态CMOS的另一个诱人之处是它在稳态工作状态时几乎完全没有功耗。正是由于同时具有稳定性和低静态功耗使静态CMOS技术已成为大多数现代数字设计的选择。对CMOS电路功耗起支配作用的是由充电和放电电容引起的动态功耗。

5.5.1 动态功耗

由充放电电容引起的动态功耗

每当电容 C_L 通过PMOS管充电时，它的电压从0升至 V_{DD}，此时从电源吸取了一定数量的能量。该能量的一部分消耗在PMOS器件中，而其余则存放在负载电容上。在由高至低的翻转期间，这一电容被放电，于是存放的能量被消耗在NMOS管中[①]。

可以推导出这一能耗的精确结果。让我们首先考虑由低至高的翻转。先假设输入波形具有为零的上升和下降时间，或者说NMOS和PMOS器件决不会同时导通。因此可以用图5.25的等效电路来表示。在这一翻转期间从电源中取得的能量值 E_{VDD} 以及在翻转结束时在电容上存储的能量 E_C 可以通过在相应周期上对瞬时功耗积分而求得：

$$E_{VDD} = \int_0^\infty i_{VDD}(t)V_{DD}dt = V_{DD}\int_0^\infty C_L\frac{dv_{out}}{dt}dt = C_LV_{DD}\int_0^{V_{DD}} dv_{out} = C_LV_{DD}^2 \qquad (5.40)$$

和

$$E_C = \int_0^\infty i_{VDD}(t)v_{out}dt = \int_0^\infty C_L\frac{dv_{out}}{dt}v_{out}dt = C_L\int_0^{V_{DD}} v_{out}dv_{out} = \frac{C_LV_{DD}^2}{2} \qquad (5.41)$$

图5.26画出了 $v_{out}(t)$ 和 $i_{VDD}(t)$ 的相应波形。

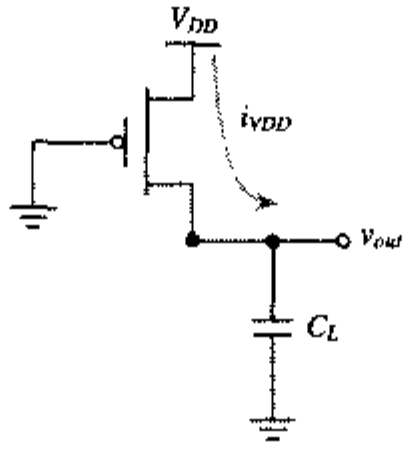


图 5.25 由低至高翻转期间的等效电路

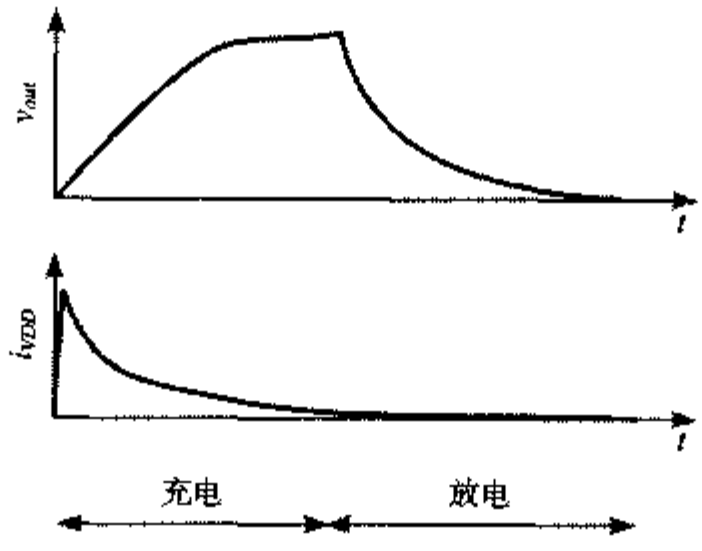


图 5.26 在 C_L 充（放）电期间的输出电压和电源电流

① 注意，这一模型是实际电路的简化。事实上，负载电容包含有多个部分，其中一些位于输出节点和GND之间，另一些则在输出节点与 V_{DD} 之间。后者经历的充电-放电周期与连接GND的电容反相位（即它们在 V_{out} 下降时充电而在 V_{out} 上升时放电）。虽然电源提供的能量分配在这两部分之间，但这并不影响总的能耗，所以本节所提供的结果仍然成立。

这些结果也可以通过观察得出：在由低至高翻转期间 C_L 被充以电荷 $C_L V_{DD}$。提供这些电荷需要从电源得到等于 $C_L V_{DD}^2 (= Q \times V_{DD})$ 的能量。存放在电容中的能量等于 $C_L V_{DD}^2/2$。这意味着由电源提供的能量只有一半是存放在 C_L 上的，另一半则由 PMOS 管消耗了。注意，这一能耗与 PMOS 器件的尺寸（因而也与电阻）无关！在放电阶段，电荷从电容上移去，因此它的能量消耗在 NMOS 器件中。同样，这一能耗与 NMOS 器件的尺寸无关。总之，每一个开关周期（由 L→H 和 H→L 翻转组成）都需要一个固定数量的能量，即 $C_L V_{DD}^2$。为了计算功耗，我们必须考虑器件的开关频率。如果这个门每秒通断 $f_{0\to1}$ 次，则功耗等于：

$$P_{dyn} = C_L V_{DD}^2 f_{0\to1} \tag{5.42}$$

式中，$f_{0\to1}$ 代表消耗能量的翻转的频率（对于静态 CMOS 为 0→1 翻转）。

工艺的进步使 $f_{0\to1}$ 不断提高（随着 t_p 的缩小）。与此同时，随着越来越多的门放在单片上，芯片上的总电容（C_L）也在增加。例如，考虑一个 0.25 μm 的 CMOS 芯片，时钟频率为 500 MHz，平均负载电容 15 fF/门，假设扇出数为 4。于是对于一个 2.5 V 的电源，每个门的功耗约等于 50 μW。对于一个 100 万门的设计，假设在每个时钟边沿处发生一次翻转，这将导致 50 W 的功耗！幸运的是，这一估计是一个悲观的估计。实际上在整片 IC 中并不是所有的门都是以 500 MHz 的全速率来开关的。在电路中实际的活动性要小得多。

例 5.11　反相器的电容功耗

现在可以很容易地计算例 5.4 中 CMOS 反相器的电容功耗。在表 5.2 中负载电容值已确定为等于 6 fF。对于 2.5 V 的电源电压，该电容充电和放电所需要的能量等于：

$$E_{dyn} = C_L V_{DD}^2 = 37.5 \text{ fJ}$$

假定该反相器以（假设的）最大可能的速率开关（$T = 1/f = t_{pLH} + t_{pHL} = 2\,t_p$）。当 t_p 为 32.5 ps（见例 5.5）时求得该电路的动态功耗为：

$$P_{dyn} = E_{dyn}/(2t_p) = 580 \text{ μW}$$

当然，在实际电路中一个反相器很少会以这一最高速率来开关，即便是，它的输出也不是在两条电源轨线电压之间摆动。因此其功耗也很低。当速率为 4 GHz（$T = 250$ ps）时，功耗降为 150 μW。这已为模拟所证实，模拟得到的功耗为 155 μW。

用 $f_{0\to1}$ 因子（也称为开关活动性，*switching activity*）来计算一个复杂电路的功耗是很麻烦的。虽然一个反相器的开关活动性很容易计算，但在比较复杂的门或电路中这就变得复杂多了。一方面一个电路的开关活动性与输入信号的本质及统计特性有关：如果输入信号保持不变，则不会发生任何开关，于是动态功耗为零！反之迅速变化的信号会引起许多次开关及功耗。其他影响开关活动性的因素有整个电路的拓扑结构以及要实现的功能。下式考虑了这些因素：

$$P_{dyn} = C_L V_{DD}^2 f_{0\to1} = C_L V_{DD}^2 P_{0\to1} f = C_{EFF} V_{DD}^2 f \tag{5.43}$$

式中，f代表输入发生变化事件的最大可能的速率（它常常就是时钟速率），而$P_{0\to1}$是时钟变化事件在该门的输出端引起0→1（即消耗功率）变化事件的概率。$C_{EFF}=P_{0\to1}C_L$称为等效电容，它代表了每时钟周期发生开关的平均电容。在我们的例子中，开关因子为10%（即$P_{0\to1}=0.1$）时使平均功耗降低至5 W。

例 5.12　开关活动性

考虑图5.27的波形。图中上面一个波形代表速想的时钟信号，而下面一个为该门输出端的信号。消耗功率的翻转每8个时钟发生两次，这相当于翻转概率为0.25（即25%）。

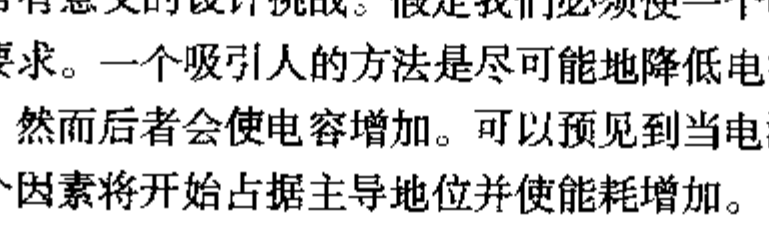


图 5.27　时钟与信号波形

低能量-功耗设计技术

随着数字集成电路日益复杂，可以预料功耗问题在未来的工艺中将会更严重。这是较低的电源电压正在变得越来越吸引人的原因之一。降低V_{DD}对P_{dyn}的影响呈二次方关系。例如在我们的例子中，使V_{DD}从2.5 V降至1.25 V将使功耗从5 W降至1.25 W。这里假设可以维持相同的时钟速率。图5.17表明只要电源电压比阈值电压高许多，这一假设并不是不现实的。但一旦V_{DD}接近$2V_T$时就会严重降低性能。

当电源电压的下限取决于外部限制（如经常在实际设计中发生的那样）或者当减小电源电压引起的性能降低不能被接受时，减少功耗的惟一方法就是减少等效电容。这可以通过减少它的两个方面来实现：实际电容及翻转活动性。

减少翻转活动性只能在逻辑和结构的抽象层次上实现，它将在第11章中比较详细地讨论。减少实际电容总体来说很值得，因为它同时也帮助改善电路的性能。由于在一个组合逻辑电路中大部分的电容是晶体管电容（栅电容和扩散电容），因此在进行低功耗设计时保持这部分电容最小是有意义的。这意味着应当保持晶体管有尽可能或合理的最小尺寸。这无疑会影响电路性能，但这一影响可以通过逻辑或结构上的加速技术来弥补。晶体管尺寸应当放大的惟一情形是当负载电容由外部电容（如扇出或导线电容）占主导地位的时候。这不同于通常单元库采用的设计方法，在单元库设计中一般都使晶体管较大以满足一定范围的负载和性能要求。

这些考虑引起了非常有意义的设计挑战。假定我们必须使一个电路的能耗最少而又同时满足所规定的对性能的最低要求。一个吸引人的方法是尽可能地降低电源电压，而用加大晶体管的尺寸来补偿性能上的损失。然而后者会使电容增加。可以预见到当电源电压足够低时，随着电源电压的进一步降低，后一个因素将开始占据主导地位并使能耗增加。■

例 5.13　确定晶体管尺寸使能耗最小

为了分析使能耗最小时确定晶体管尺寸的问题，我们考察一个静态CMOS反相器驱动一个外

部负载电容 C_{ext} 时的简单情形，如图 5.28 所示。为了考虑输入负载效应，假设反相器本身由一个最小尺寸器件驱动。我们的目的是使整个电路的能耗最小而又保持最低的性能要求。设计的自由度是该电路反相器的尺寸系数 f 和电源电压 V_{dd}。优化后电路的传输延时应当不大于参数为 $f = 1$ 及 $V_{dd} = V_{ref}$ 的参考电路的传播延时。

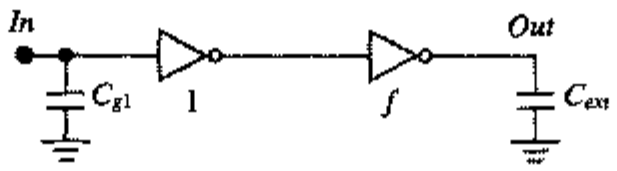


图 5.28　CMOS 反相器驱动一个外部负载电容 C_{ext}，而它本身又由一个最小尺寸的门驱动

利用 5.4.3 节介绍的方法（确定反相器链的尺寸），我们可以推导出该电路传播延时的表达式如下：

$$t_p = t_{p0}\left(\left(1+\frac{f}{\gamma}\right)+\left(1+\frac{F}{f\gamma}\right)\right) \tag{5.44}$$

式中，$F = C_{ext}/C_{g1}$ 是该电路总的等效扇出，而 t_{p0} 是反相器的本征延时。它与 V_{DD} 的关系可以用由公式（5.21）推导出的下列表达式来近似：

$$t_{p0} \sim \frac{V_{DD}}{V_{DD}-V_{TE}} \tag{5.45}$$

其中，$V_{TE} = V_T + V_{DSAT}/2$。一旦电路的总电容已知，则在输入端单个翻转的能耗很容易计算出：

$$E = V_{dd}^2 C_{g1}((1+\gamma)(1+f)+F) \tag{5.46}$$

现在，性能约束就是指尺寸放大电路的传播延时应当等于（或小于）参考电路（$f = 1$，$V_{dd} = V_{ref}$）的延时。为了简化以下的分析，我们假设该门的本征输出电容等于它的栅电容，即 $\gamma = 1$。因此，

$$\frac{t_p}{t_{pref}} = \frac{t_{p0}\left(2+f+\frac{F}{f}\right)}{t_{p0ref}(3+F)} = \left(\frac{V_{DD}}{V_{ref}}\right)\left(\frac{V_{ref}-V_{TE}}{V_{DD}-V_{TE}}\right)\left(\frac{2+f+\frac{F}{f}}{3+F}\right) \approx 1 \tag{5.47}$$

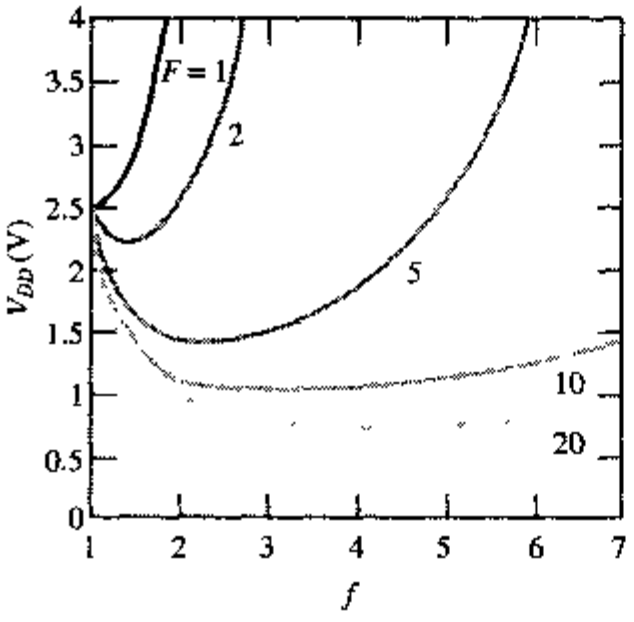
公式（5.47）建立了尺寸系数 f 和电源电压之间的关系，图 5.29（a）画出了对于不同 F 值时的这一关系。这些曲线都有一个明显的最小值。由最小尺寸起增加反相器的尺寸最初会使性能提高，因此允许降低电源电压。这在达到最优尺寸系数 $f = \sqrt{F}$ 之前一直都是有效的，对于仔细阅读过前面几节的读者而言应当不会对此感到任何意外。进一步加大器件尺寸只会增加自载系数而降低性能，因此需要提高电源电压。同时注意到对于 $F = 1$，参考电路的情形是最好的结果，尺寸的任何增加只会加大自载影响。

有了 $V_{DD}(f)$ 的关系，就可以推导尺寸放大电路的能量（归一至参考电路）与尺寸系数 f 之间的关系：

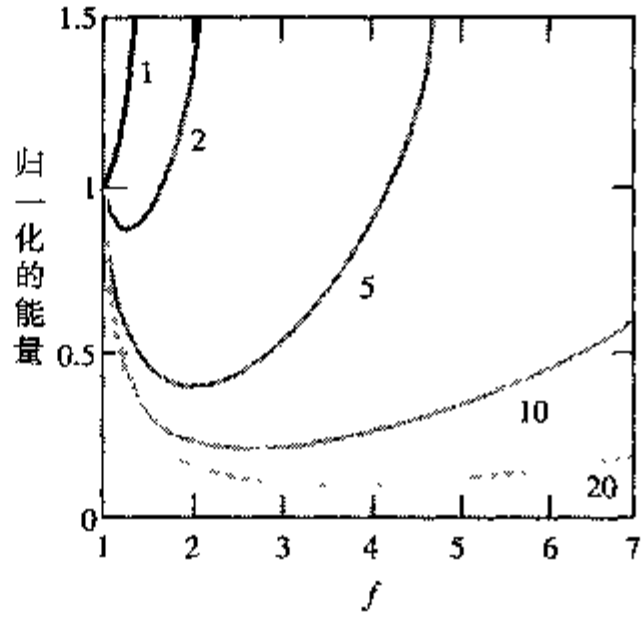
$$\frac{E}{E_{ref}} = \left(\frac{V_{DD}}{V_{ref}}\right)^2\left(\frac{2+2f+F}{4+F}\right) \tag{5.48}$$

推导出最优尺寸系数的解析表达式是可能的，但这会得到非常复杂和凌乱的公式。而图解法就很

有效。所得到的图画在图5.29（b）中，由此可以得出以下几个结论[1]：



（a）对总等效扇出 F 的不同值所要求的电源电压与尺寸系数 f 的关系



（b）放大了尺寸后电路的能量（归一至参考值的情形）与 f 的关系。$V_{ref}=2.5$ V，$V_{TE}=0.5$ V

图 5.29 最小化能耗时反相器尺寸的确定

- 改变器件尺寸并降低电源电压是减小一个逻辑电路能耗的非常有效的方法。对于具有较大等效扇出的电路尤为如此，因为在这些电路中可以达到几乎十倍的能量减少。这一收益对于较小的 F 值是相当大的。惟一的例外是 $F=1$ 的情形，此时最小尺寸的器件也是最有效的器件。
- 在最优值之外过多地加大晶体管的尺寸会付出较大的能量代价。但遗憾的是，这在今天的许多设计中是普遍采用的一种方法。
- 考虑能量时的最优尺寸系数小于考虑性能时的最优尺寸系数，在 F 值较大时尤其如此。例如当扇出为20时，f_{opt}（能量）= 3.53，而 f_{opt}（性能）= 4.47。一旦 V_{DD} 开始接近 V_{TE}，加大器件尺寸只能很少地降低电源电压，因此能耗的降低也非常少。

直接通路电流引起的功耗

在实际设计中，假设输入波形的上升和下降时间为零是不正确的。输入信号不为无穷大的斜率造成了开关过程中 V_{DD} 和 GND 之间在短期内出现一条直流通路，此时 NMOS 和 PMOS 管同时导通。这一情形显示在图 5.30 中。在（合理）假设所形成的电流脉冲可近似成三角形及反相器的上升和下降响应是对称的条件下，可以计算出每个开关周期消耗的能量如下：

$$E_{dp} = V_{DD}\frac{I_{peak}t_{sc}}{2} + V_{DD}\frac{I_{peak}t_{sc}}{2} = t_{sc}V_{DD}I_{peak} \qquad (5.49)$$

计算平均功耗为：

$$P_{dp} = t_{sc}V_{DD}I_{peak}f = C_{sc}V_{DD}^2 f \qquad (5.50)$$

直接通路引起的功耗与开关活动性（switching activity）成正比，这类似于电容功耗。t_{sc} 代表两个器件同时导通的时间。对于一个直线输入斜率，可以用公式（5.51）求得它的近似值，式中 t_s 代表0~100%的翻转时间。

① 在第 11 章中将从更广的意义上再次谈及其中的一些结论。

$$t_{sc} = \frac{V_{DD}-2V_T}{V_{DD}}t_s \approx \frac{V_{DD}-2V_T}{V_{DD}} \times \frac{t_{r(f)}}{0.8} \qquad (5.51)$$

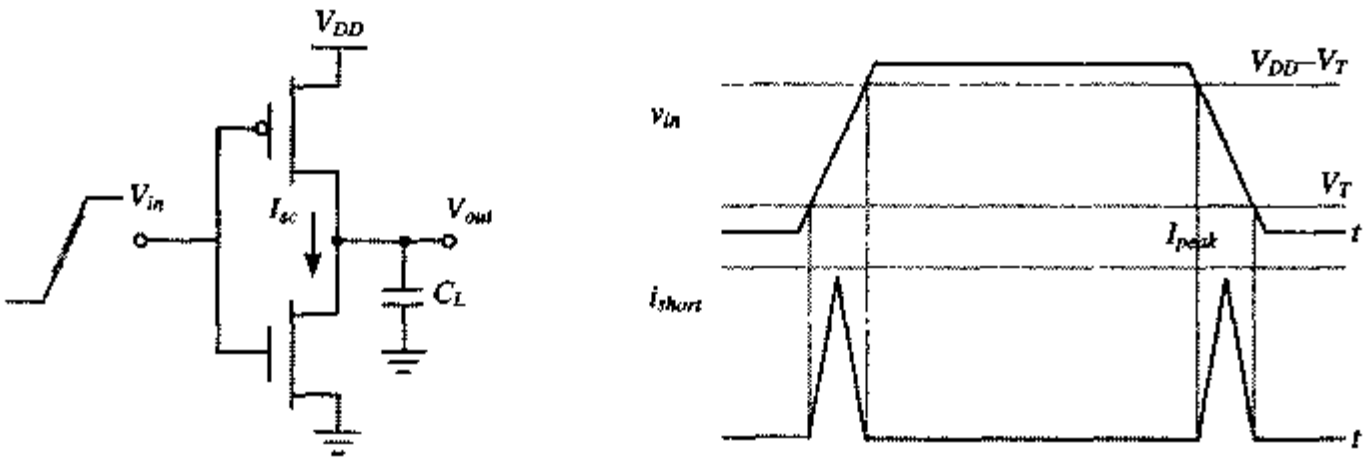
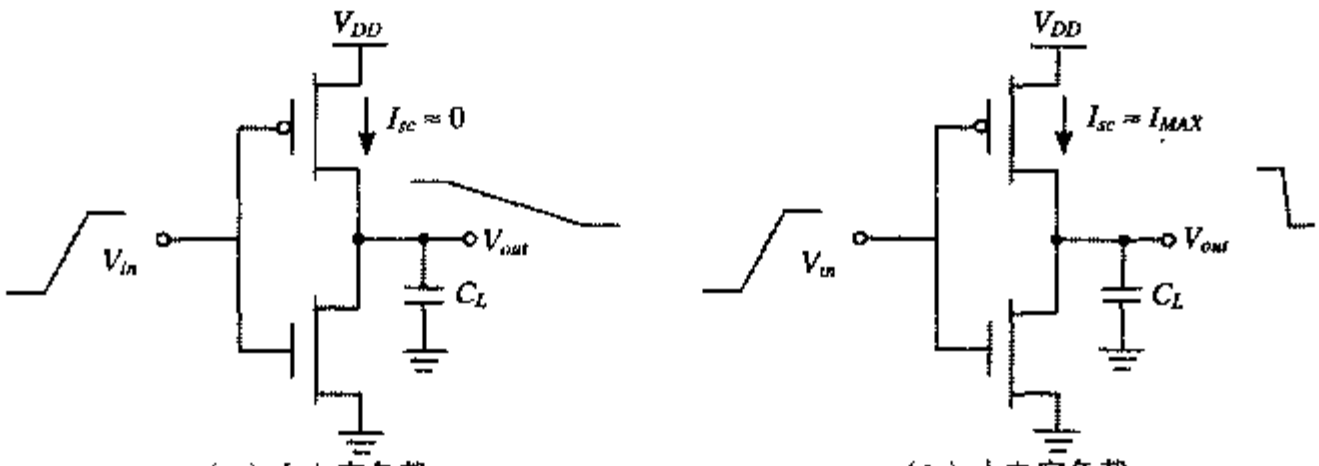


图 5.30　过渡期间的短路电流

I_{peak} 由器件的饱和电流决定，因此直接正比于晶体管的尺寸。峰值电流也与输入和输出斜率之比密切相关。这一关系可用以下简单的分析得到最好的说明：考虑一个静态 CMOS 反相器在输入端发生由 0→1 的翻转。首先假设负载电容很大，所以输出的下降时间明显大于输入的上升时间［见图 5.31（a）］。在这些情况下输入在输出开始改变之前就已经通过了过渡区。由于在这一时期 PMOS 器件的源–漏电压近似为 0，因此该器件甚至还没有传导任何电流就断开了。在这种情况下短路电流接近于零。现在考虑相反的情况，就是输出电容非常小，因此输出的下降时间明显小于输入的上升时间［见图 5.31（b）］。PMOS 器件的源–漏电压在翻转期间的大部分时间内等于 V_{DD}，从而引起了最大的短路电流（等于 PMOS 的饱和电流）。这显然代表了最坏情况的条件。以上分析的结论在图 5.32 中得到证实，图中画出了在由低至高翻转期间通过 NMOS 管的短路电流与负载电容的关系。



（a）大电容负载　　（b）小电容负载

图 5.31　负载电容对短路电流的影响

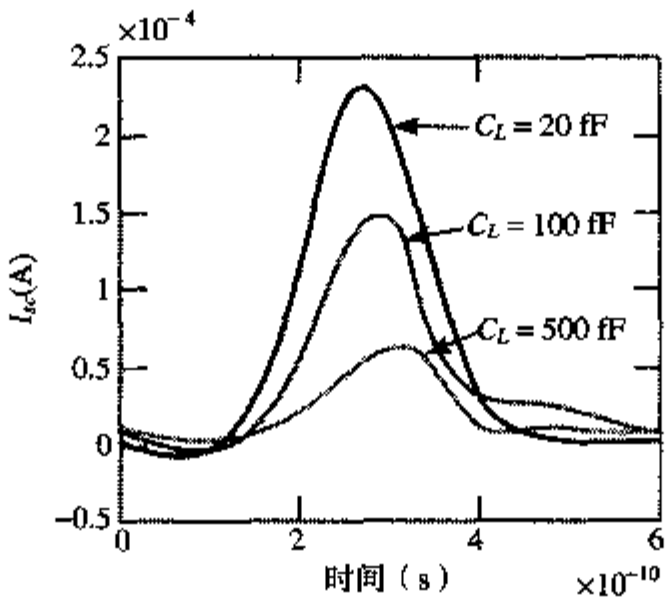


图 5.32　CMOS 反相器通过 NMOS 晶体管的短路电流与负载电容的关系（输入斜率固定为 500 ps）

这一分析的结论是：使输出的上升/下降时间大于输入的上升/下降时间可以使短路功耗减到最小。但输出的上升/下降时间太大会降低电路的速度并在扇出门中引起短路电流。这个例子很好地说明了只顾局部优化而不管全局是如何会引起不良后果的。

设计技术

一个更为实用的从全局角度优化功耗的规则可以正式说明如下（[Veendrick84]）：

短路电流功耗可以通过使输入和输出信号的上升/下降时间匹配来达到最小。在整个电路层次上，这意味着所有信号的上升/下降时间应当保持在一定范围内不变。

使一个门的输入和输出上升时间相等就这个具体的门本身而言并不是一个最优的结果，但却能保持整个短路电流在界定的范围内。这显示在图5.33 中，图中画出了一个反相器的短路能耗（归一于零输入上升时间的能耗）与输入和输出上升/下降时间之比 r 之间的关系。对于一个给定的反相器尺寸（对 V_{DD}=5 V，r >2...3）当负载电容太小时，功耗主要来自短路电流。对于非常大的负载电容值，所有的功耗都用来充电和放电负载电容。如果使输入和输出的上升/下降时间相等，则大部分功耗与动态功耗有关，而只有很小一部分（<10%）出自短路电流。

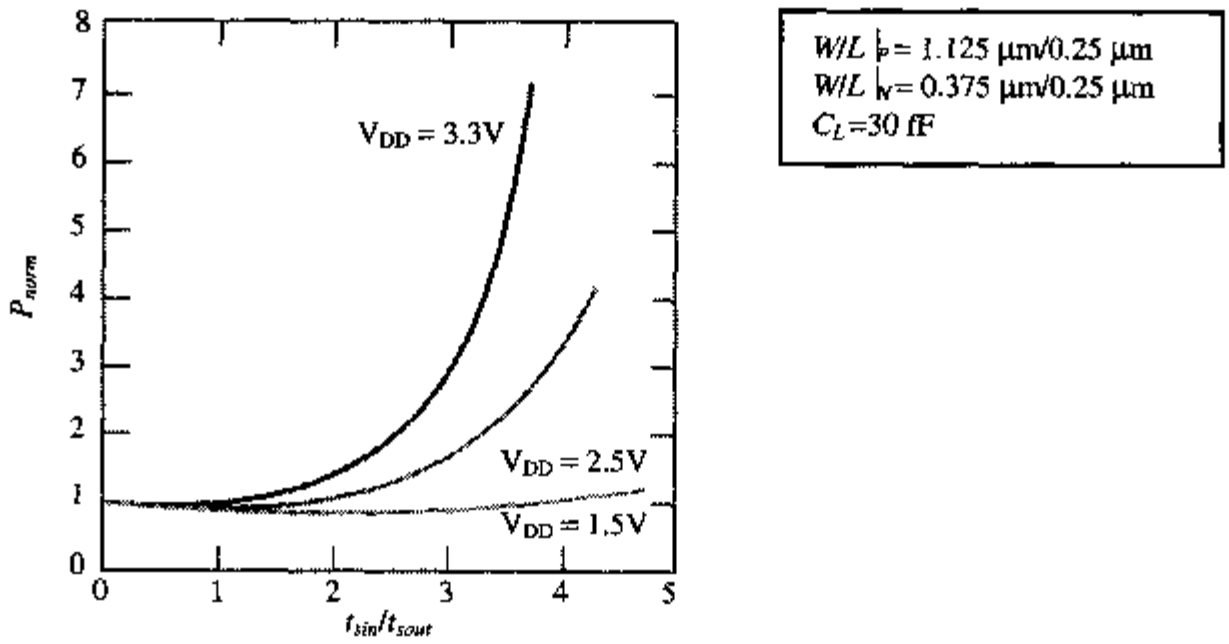


图 5.33 一个静态 CMOS 反相器的功耗与输入和输出上升/下降时间之比的关系。功耗归一至零输入上升时间时的功耗。当斜率（上升/下降时间）之比较小时，输入/输出间的耦合将引起某些额外的功耗

我们还注意到当降低电源电压时短路电流的影响减小，这从公式（5.51）中可以清楚地看到。在极端情形下当 $V_{DD}<V_{Tn}+|V_{Tp}|$ 时，短路功耗完全消除，因为两个器件决不会同时导通。当阈值电压以比电源电压低的速率下降时，短路功耗在深亚微米工艺中将变得较不重要。当电源电压为2.5 V 及阈值为0.5 V 左右时，要求输入/输出斜率比为2 才能使功耗下降10%。 ■

最后，值得注意的是，短路功耗可以通过增加一个负载电容 $C_{sc}=t_{sc}\,I_{peak}/V_{DD}$ 与 C_L 并联来模拟，这在公式（5.50）中看得很清楚。这一短路电容值与 V_{DD}、晶体管的尺寸以及输入/输出斜率比有关。

5.5.2 静态功耗

一个电路的静态（或称稳态）功耗可用下列关系来表示：

$$P_{stat} = I_{stat}V_{DD} \tag{5.52}$$

式中，I_{stat} 是在没有开关活动存在时在电源两条轨线之间流动的电流。

理想情况是 CMOS 反相器的静态电流为零，因为 PMOS 和 NMOS 器件在稳态工作状况下决不

会同时导通。可惜的是总会有泄漏电流流过位于晶体管源（或漏）与衬底之间的反相偏置的二极管结，如图5.34所示。这一电流一般来说是非常小的，因此可以被忽略。对于所考虑的器件尺寸，在室温下每单位漏极面积的泄漏电流在 10~100 pA/μm^2 之间。对于一个含 100 万门的芯片来说，若每个门的漏极面积为0.5 μm^2 并在2.5 V 的电源电压下工作，则在最坏情况下因二极管漏电引起的功耗等于0.125 mW，显然，这不是什么严重问题。

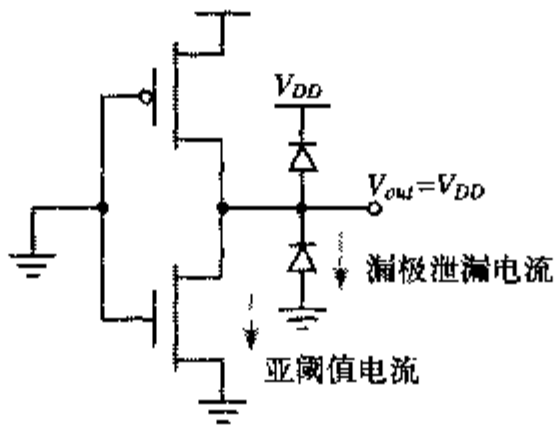


图 5.34　CMOS 反相器中泄漏电流的来源（$V_{in}=0$ V）

然而，应当知道结的泄漏电流是由热产生的载流子引起的。它们的数值随结温而增加，并且呈指数关系。在85°C（民用硬件通常规定的结温上限）时，泄漏电流为室温时的60倍。因此保持一个电路总的工作温度较低是所期望的目标。由于温度与消耗的热及散热机理有很大的关系，因此要达到这一目的只能通过限制电路的功耗或使用能支持有效散热的封装。

泄漏电流的一个越来越突出的来源是晶体管的亚阈值电流。正如在第 3 章中讨论过的，一个 MOS 管甚至在 V_{GS} 小于阈值电压时也可以有一个漏-源电流（见图5.35）。阈值电压越是接近 0 V，则在 $V_{GS}=0$ V 时的泄漏电流越大，因面静态功耗也就越大。为了抵消这一效应，器件的阈值电压一般应当保持足够高。标准工艺的特征值 V_T 从未小于 0.5~0.6 V，有时甚至还相当大（~0.75 V）。

随亚微米工艺尺寸的缩小，同时出现了电源电压降低，因而这一方法受到挑战（这在图 3.41 中已非常明显）。我们在前面已经总结过（见图 5.17），降低电源电压同时保持阈值电压不变会造成性能的严重损失，特别是当 V_{DD} 接近于 $2V_T$ 时。解决这一性能问题的一个方法是同时降低这一器件的阈值电压。这使图5.17 中的曲线左移，它意味着由降低电源电压造成的性能损失减小。可惜的是，阈值电压的最低值是由所允许的亚阈值漏电流的数量所决定的，如图5.35 所示。因此阈值电压的选择代表了在性能和静态功耗之间的权衡取舍。电源电压的继续降低预示着新一代 CMOS 工艺的出现，但它也迫使阈值电压更为降低，从而使亚阈值导电成为功耗的主要来源。因此能生产具有迅速彻底关断特性的器件的工艺技术将变得更加引人注目。后者的一个例子是绝缘体上硅（SOI，Silicon-on-Insulator）技术，它的 MOS 管具有一个接近理想 60 mV/十倍电流的斜率系数。

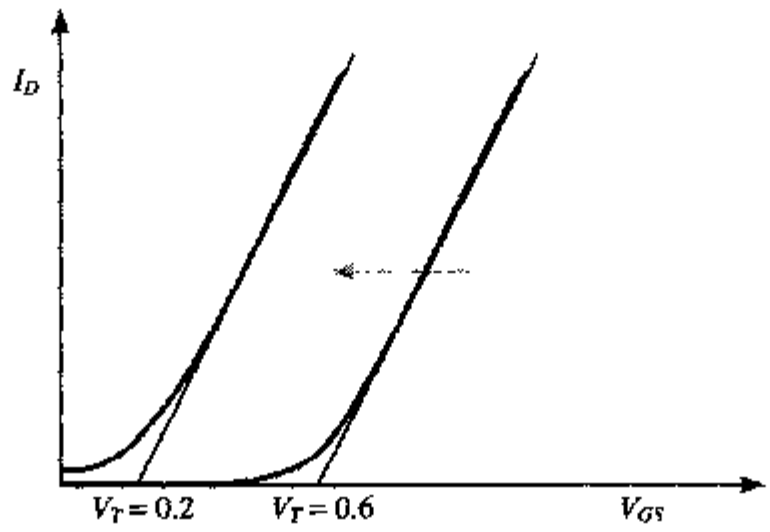


图 5.35　$V_{GS}=0$ 时降低阈值会使亚阈值电流增加

例 5.14 阈值降低对器件性能和静态功耗的影响

考虑0.25 μm CMOS工艺的一个最小尺寸NMOS晶体管。在第3章中我们已推导出该器件的斜率系数S等于90 mV/十倍电流。当V_T约为0.5 V时晶体管的关断状态电流（在$V_{GS}=0$时）等于10^{-11}A（见图3.22）。使阈值降低200 mV达到0.3 V时，晶体管的关断状态电流加大170倍！假设一个100万门设计的电源电压为1.5 V，这意味着静态功耗为$10^6 \times 170 \times 10^{-11} \times 1.5 = 2.6$ mW。进一步降低阈值至100 mV时所产生的功耗几乎接近0.5 W，这是无法接受的！但在这一电源电压下阈值的降低相当于性能分别提高了25%和40%。

阈值的下限在某种意义上是人为决定的。一个静态CMOS电路的泄漏电流必须为零的概念是不正确的。无疑，漏电流的存在会减小噪声容限，因为逻辑电平不再等于全部电源电压，但只要噪声容限在一定范围之内，这不算什么严重问题。自然，泄漏电流使静态功耗增加。但它可以用降低电源电压来补偿，这又可以不降低电路的性能而通过降低阈值电压来达到，结果使动态功耗以平方关系下降。对于0.25 μm CMOS工艺，以下两组电路配置得到相同的性能：3 V电源-V_T 0.7 V；0.45 V电源-V_T 0.1 V。但是后者的动态功耗要小45倍[Liu93]！因此选择正确的电源值和阈值电压值需要再次权衡利弊。最佳工作点取决于电路的活动性。当存在相当大的静态功耗时，非常重要的是使不工作（nonactive）的模块暂时断电（*powered down*）以免静态功耗成为支配因素。暂时断电（亦称待机状态）可以通过切断该工作单元与电源线的连接或降低电源电压来完成。

5.5.3 综合考虑

CMOS反相器的总功耗现在可以表示成三部分的和：

$$P_{tot} = P_{dyn} + P_{dp} + P_{stat} = (C_L V_{DD}^2 + V_{DD} I_{peak} t_s) f_{0\to 1} + V_{DD} I_{leak} \qquad (5.53)$$

在典型的CMOS电路中电容功耗是占主导地位的因素。直接通路功耗可以通过细心的设计控制在限定范围之内，因此不应当成为问题。漏电目前可以忽略，但在不久的将来这种情形会有所改变。

功耗-延时积或每操作的能量损耗

在第1章已经介绍了可将功耗-延时积（PDP）作为一个逻辑门的质量评定指标：

$$PDP = P_{av} t_p \qquad (5.54)$$

*PDP*是能量的衡量，这从它的单位（W × s = 焦耳）就可以清楚地看出。假设这个门以其最大可能的速率$f_{max} = 1/(2t_p)$切换，并忽略静态和直接通路电流引起的功耗，我们得到：

$$PDP = C_L V_{DD}^2 f_{max} t_p = \frac{C_L V_{DD}^2}{2} \qquad (5.55)$$

这里，*PDP*代表每次开关（即0→1或1→0的翻转）消耗的能量。我们通常感兴趣的是每个翻转周期所消耗的能量E_{av}。由于每个反相器周期包括一个0→1和一个1→0的翻转，因此E_{av}是*PDP*的两倍。

能量-延时积

用*PDP*作为衡量工艺技术或逻辑门拓扑结构的质量指标是有问题的。它衡量了开关这个门所需要的能量，这是一个重要的特性。但是对于一个给定的结构这个数字可以通过降低电源电压而任意缩小。从这一角度来看，使这个电路工作的最优电压应当是仍能保证功能的最低可能的电压

值。但正如前面已讨论过的，这要以牺牲性能为代价。一个更合适的指标应当把性能和能量的度量放在一起考虑。能量-延时积（或称 *EDP*）就是这样一个指标：

$$EDP = PDP \times t_p = P_{av} t_p^2 = \frac{C_L V_{DD}^2}{2} t_p \tag{5.56}$$

值得分析一下EDP与电压的关系。较高的电源电压能够减少延时，但会增加能耗，电压低时则正好相反。因此应当存在一个最优工作点。假设NMOS和PMOS管具有可比拟的阈值电压和饱和电压，我们可以把传播延时公式（5.21）简化为：

$$t_p \approx \frac{\alpha C_L V_{DD}}{V_{DD} - V_{TE}} \tag{5.57}$$

式中，$V_{TE} = V_T + V_{DSAT}/2$，α为工艺参数。联立公式（5.56）和公式（5.57）[①]得到：

$$EDP = \frac{\alpha C_L^2 V_{DD}^3}{2(V_{DD} - V_{TE})} \tag{5.58}$$

在公式（5.58）中对 V_{DD} 求导并令结果为零，即得到最优电源电压。结果为：

$$V_{DDopt} = \frac{3}{2} V_{TE} \tag{5.59}$$

由这一分析得到了一个能同时优化性能和能耗的较低的电源电压值。对于阈值在 0.5 V 范围的亚微米工艺最优电源电压为1 V 左右。

例 5.15　0.25 μm CMOS 反相器的最优电源电压

从第 3 章列出的通用CMOS工艺的工艺参数中可以推导出 V_{TE} 值如下：

$$V_{Tn} = 0.43\ \text{V},\ V_{Dsatn} = 0.63\ \text{V},\ V_{TEn} = 0.74\ \text{V}$$
$$V_{Tp} = -0.4\ \text{V},\ V_{Dsatp} = -1\ \text{V},\ V_{TEp} = -0.9\ \text{V}$$
$$V_{TE} \approx (V_{TEn} + |V_{TEp}|)/2 = 0.8\ \text{V}$$

因此，$V_{DDopt} = (3/2) \times 0.8\ \text{V} = 1.2\ \text{V}$。图 5.36 的模拟结果画出了归一化的延时、能量以及能量-延时积，它证实了这一结果。所预测的最优电源电压为 1.1 V。该图清楚地显示了在延时与能量之间的互换关系。

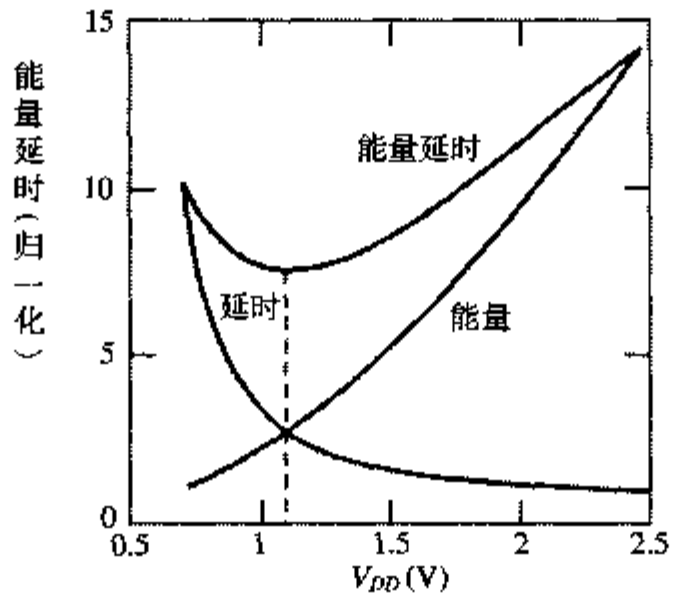


图 5.36　0.25 μm CMOS 反相器归一化的延时、能量及能量-延时积的曲线图

① 该公式仅在器件保持速度饱和时是精确的，这一点在电源电压较低时可能达不到。这会给分析带来一些不精确性，但不会影响整体结果。

警告：上面的例子虽然显示了存在一个使门的能量-延时积最小的电源电压，但这一电压对于一个给定的设计并不一定就代表最优电压。例如某些设计要求延时最小的性能，这需要以能量为代价的较高电压。同样，一个低能量设计可以通过工作在低电压下并采用像流水线或并行性这样的结构技术获取总的系统性能来实现。

5.5.4 利用 SPICE 分析功耗

在第1章已经定义了一个电路的平均功耗，为方便起见在此重复一下：

$$P_{av} = \frac{1}{T}\int_0^T p(t)dt = \frac{V_{DD}}{T}\int_0^T i_{DD}(t)dt \qquad (5.60)$$

式中，T 是相关的周期，V_{DD} 和 i_{DD} 分别为电源电压和电流。某些SPICE实现的版本含有一些功能可以计算一个电路信号的平均值。例如 HSPICE.MEASURE TRAN I(VDD) AVG 命令可以计算在已算出的瞬态响应曲线（I(VDD)）下的面积并把它除以相关的周期。这一结果与公式（5.60）给出的定义完全一样。可惜的是，SPICE 的其他实现版本功能并没有这么强。但只要我们认识到SPICE实际上就是一个微分方程的求解工具，那么这一点并不像看上去那么差。我们可以很容易地想像出一个简单的电路作为一个积分器，它的输出信号就是平均功率。

例如，考虑图 5.37 的电路。电源送出的电流由一个电流控制的电流源来度量，并且在电容 C 上进行积分。电阻 R 只是为了 DC（直流）收敛，因此应当选择得尽可能地大以使漏电流最小。明智地选择元件的参数可以保证输出电压 P_{av} 等于平均功耗。该电路的工作可以用公式（5.61）来概括，这里假设电容 C 上的初始电压为零：

$$C\frac{dP_{av}}{dt} = ki_{DD} \qquad (5.61)$$

或

$$P_{av} = \frac{k}{C}\int_0^T i_{DD}dt$$

由公式（5.60）及公式（5.61）可得到等效电路参数必须满足的条件：$k/C = V_{DD}/T$。在这些条件下，所示的等效电路就是一个可用来跟踪数字电路平均功耗的方便工具。

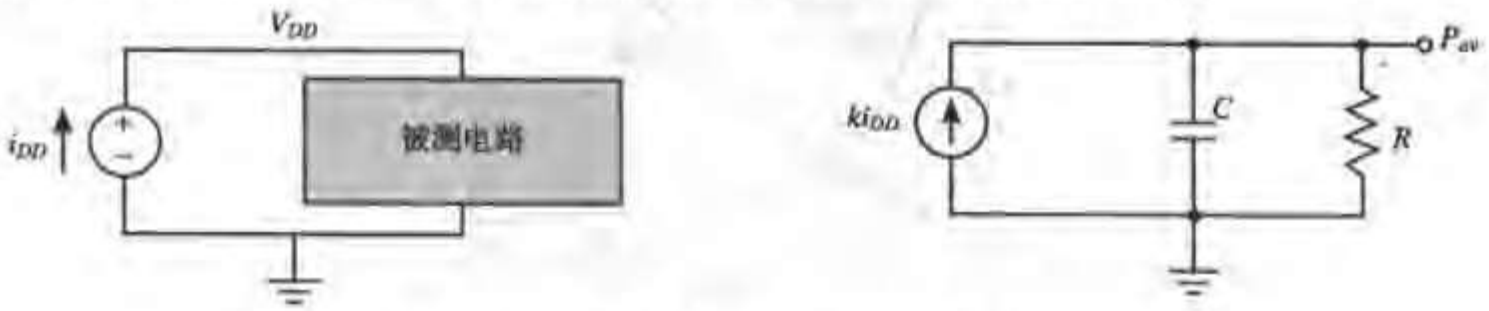


图 5.37 用 SPICE 度量平均功耗的等效电路

例 5.16 反相器的平均功耗

例5.4的反相器在250 ps翻转周期中的平均功耗可以用以上的方法来分析（T = 250 ps，k = 1，

V_{DD} = 2.5 V，因此 C = 100 pF）。所得到的功耗画在图 5.38 中，该图显示了平均功耗约为157.3 μW。而由.MEAS AVG 命令得到的值为 160.3 μW，这表明这两种方法近似等效。这些数字相当于能量为39 fJ（它接近在例5.11 中推导出的37.5 fJ）。我们注意到在由高至低翻转期间功耗稍微有点负方向的跌落。这是由于当在输入和输出间的电路耦合使输出短暂地超过 V_{DD} 时电流注入电源的缘故（如图 5.16 中的瞬态响应所示）。

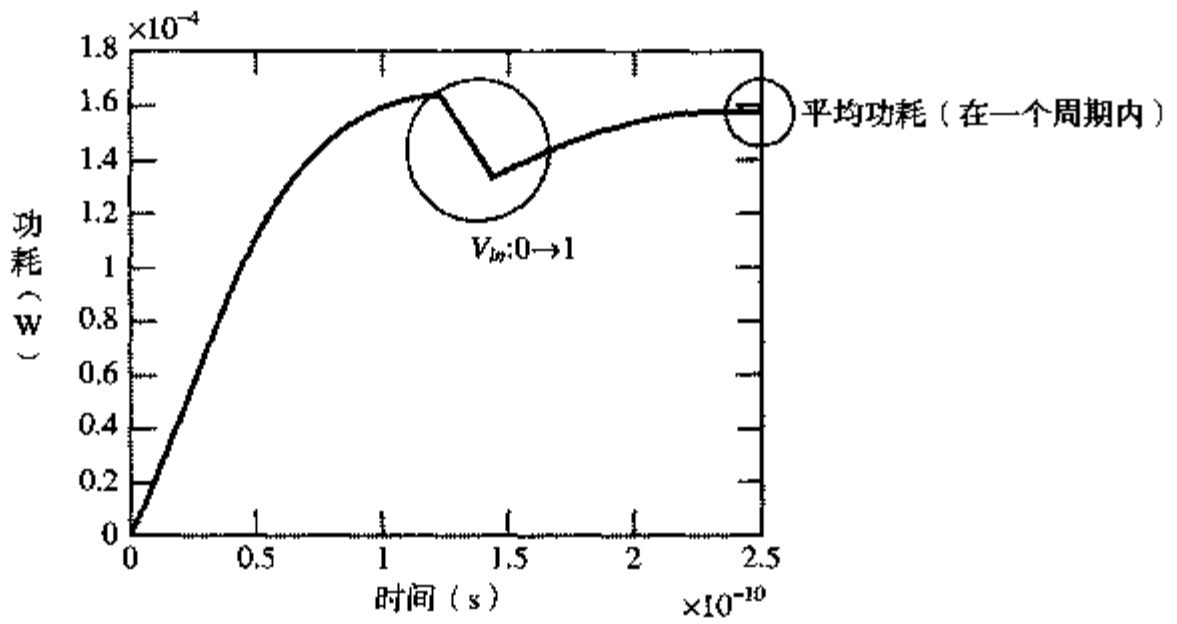


图 5.38　利用 SPICE 计算功耗

5.6　综述：工艺尺寸缩小及其对反相器衡量指标的影响

在 3.5 节中已展示了工艺尺寸缩小对一些重要设计参数的影响，如面积、延时和功耗。为了清楚起见，在这里重复列出工艺尺寸缩小表（表 3.8）中一些比较重要的条目。

这些理论预见的合理性可以通过回顾和观察过去几十年间的趋势而得到证实。由图 5.39 可以看到门延时确实以每年 13% 的速率呈指数下降，或者说每五年减半。这一下降速率与表 5.4 的预见一致，因为 S 的平均值如在图 3.40 中看到的那样大约为 1.15。一个两输入 NAND 门扇出为 4 时的延时已从 20 世纪 60 年代的几十纳秒下降到 2000 年的十分之一纳秒，并且预见到 2010 年时将为几十皮秒（1 皮秒为 10^{-12} 秒）。

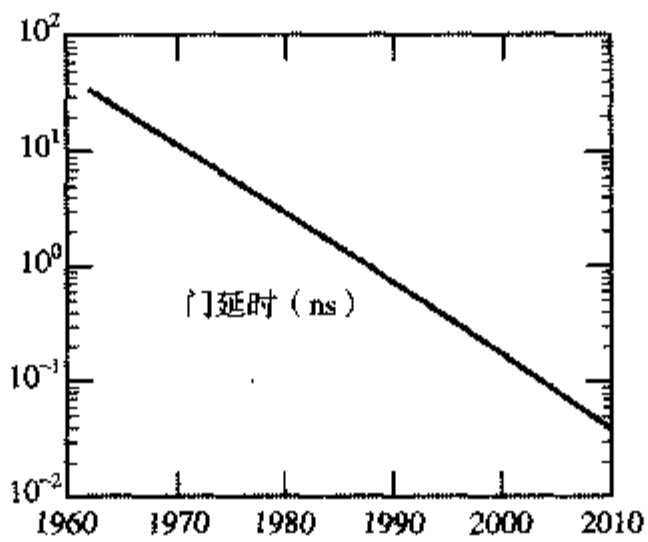


图 5.39　门延时的减小（摘自［Dally98］）

表 5.4 短沟器件的尺寸缩小情形（S 和 U 分别代表工艺和电压的缩小参数）

参数	关系	全比例缩小	一般化缩小	恒压缩小
面积/器件	W–L	$1/S^2$	$1/S^2$	$1/S^2$
本征延时	$R_{on}C_{gate}$	$1/S$	$1/S$	$1/S$
本征能量	$C_{gate}V^2$	$1/S^3$	$1/SU^2$	$1/S$
本征功率	能量/单位延时	$1/S^2$	$1/U^2$	1
功率密度	功率/单位面积	1	S^2/U^2	S^2

直到最近，降低功耗还只是第二位考虑的问题。因此对于每个门或设计的功耗只有一些统计数据。图 5.40 是一个有趣的图，它画出了对 1980 年至 1995 年期间完成的大量设计所统计的功率密度。尽管差别很大（甚至对固定的一种工艺也是如此），但它表明功率密度近似地随 S^2 增加。这相应于表 5.4 所显示的固定电压缩小的情形。在最近几年，我们可以期望功率密度下降的情形与全缩小模型更为一致——这一模型预见到的是不变的功率密度——这是由于电源电压在加速降低以及对低功耗设计技术日益关注的缘故。即使在这些情形下，每个芯片的功耗也将由于越来越大的芯片尺寸而继续增长。

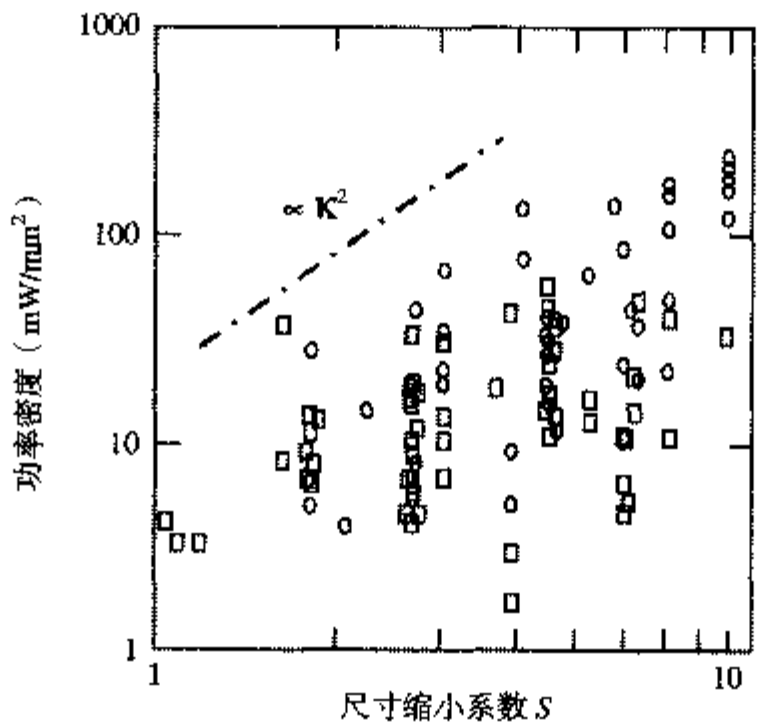


图 5.40 微处理器和数字信号（DSP）处理器中功率密度的增加与尺寸缩小系数 S 的关系（［Kuroda95］）。S 对于 4 μm 工艺归一至 1

然而以上所介绍的缩小模型有一个主要的缺陷。性能和功耗的预测只得到仅考虑器件参数的“本征”数值。在第 4 章中我们已得出结论，即互连线表现出不同的尺寸缩小特性，而且连线的寄生参数可能支配整个性能。同样，连线电容充电和放电的能量也可能占据总能量的主要部分。为了得到比较清晰的观点，我们必须建立起一个联合的模型，它同时考虑了器件和连线的缩小模型。连线电容及其尺寸缩小特性的影响总结在表 5.5 中。这里采用的是第 4 章中所介绍的固定电阻模型。我们进一步假定驱动器电阻比连线电阻更主要，对于短连线至中长连线来说情形确实如此。

这一模型预见到随着工艺尺寸的缩小，互连引起的延时（及能耗）将越来越重要。它的影响对于短连线（$S = S_L$）而言只是由于 ε_c 的增加而增加，但对于中等范围的长连线（$S_L < S$）而言将变得日益显著。这些结论已为许多研究所证实，图 5.41 为其中的一个例子。连线影响与本征部分的比今后实际上将如何变化是一个可争议的问题，因为它取决于范围很广的独立参数，如系统总体

结构、设计方法学、晶体管尺寸以及互连材料等。互连线在不久的将来很快就会使CMOS 性能“饱和”而走向末日的说法也许是过于夸大了。但十分清楚的是对互连线的日益关注是绝对必要的，并且这也许会改变下一代电路设计和优化的方式（例如 [Sylvester98]）。

表 5.5　导线电容的缩小情形。S 和 U 分别代表工艺和电压的缩小参数，而 S_L 代表导线长度变化的比例系数。ε_c 代表边缘电容和线间电容的影响

参数	关系	一般化缩小
导线电容	WL/t	ε_c/S_L
导线延时	$R_{on}C_{int}$	ε_c/S_L
导线能量	$C_{int}V^2$	ε_c/S_LU^2
导线延时/本征延时		ε_cS/S_L
导线能量/本征能量		ε_cS/S_L

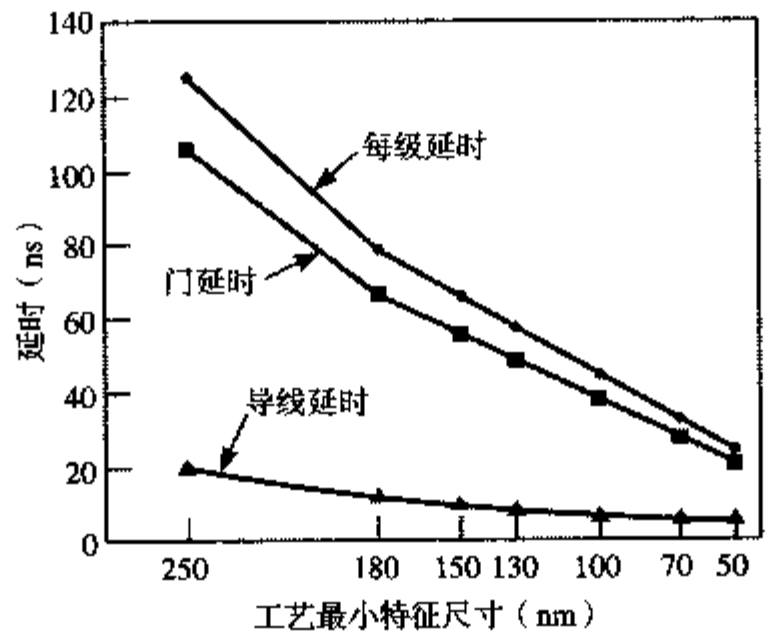


图 5.41　导线延时与门延时的比随工艺进步而变化（摘自 [Fisher98]）

5.7　小结

本章对静态CMOS 反相器进行了严格和深入的分析。该门的主要特点总结如下：

- 静态 CMOS 反相器把一个上拉的 PMOS 器件和一个下拉的 NMOS 器件组合在一起。因为 PMOS 具有较低的电流驱动能力，通常应使它比NMOS 宽。
- 该门具有几乎理想的电压传输特性。逻辑摆幅等于电源电压并且与晶体管的尺寸无关。一个对称反相器（它的PMOS 和NMOS 管具有相同的电流驱动强度）的噪声容限接近 $V_{DD}/2$。稳态响应不受扇出的影响。
- 它的传播延时主要由充放电负载电容 C_L 所需要的时间决定。作为一阶近似，它可以近似为下式：

$$t_p = 0.69C_L\left(\frac{R_{eqn}+R_{eqp}}{2}\right)$$

使负载电容保持较小是实现高性能电路的最有效手段。只要延时主要受扇出和导线等外部

（或负载）电容的影响，改变晶体管的尺寸就可能有助于提高性能。

- 功耗主要是由在充电和放电负载电容时消耗的动态功耗决定的。它为 $P_{0\to1}C_LV_{DD}{}^2f$。功耗与电路的活动性成正比。在开关通断期间发生的直接通路电流所引起的功耗可以通过对信号斜率的仔细修正来限制。静态功耗通常可以忽略，但在将来由于亚阈值电流的原因它可能成为一个主要因素。
- 使工艺尺寸变小是减少一个门的面积、传播延时以及功耗的有效手段。如果电源电压也同时降低，则其影响甚至更为惊人。
- 互连线的影响将在总延时和总性能中逐渐占有更大的比例。

5.8 进一步探讨

CMOS 反相器的工作已是许多出版物和教科书的内容。实际上每一本关于数字设计的书都使用了大量的篇幅来分析基本反相器门。在第 1 章中已经列出了广泛的参考资料。以下是本章中曾引用过的一些特别有意义的参考资料。

参考文献

[Dally98] W. Dally and J. Poulton, *Digital Systems Engineering*, Cambridge University Press, 1998.

[Fisher98] P. D. Fisher and R. Nesbitt, "The Test of Time: Clock-Cycle Estimation and Test Challenges for Future Microprocessors," *IEEE Circuits and Devices Magazine*, 14(2), pp. 37–44, 1998.

[Hedenstierna87] N. Hedenstierna and K. Jeppson, "CMOS Circuit Speed and Buffer Optimization," *IEEE Transactions on CAD*, vol. CAD-6, no. 2, pp. 270–281, March 1987.

[Kuroda95] T. Kuroda and T. Sakurai, "Overview of low-power ULSI circuit techniques," *IEICE Trans. on Electronics*, vol. E78-C, no. 4, pp. 334-344, April 1995.

[Liu93] D. Liu and C. Svensson, "Trading speed for low power by choice of supply and threshold voltages," *IEEE Journal of Solid-State Circuits*, vol. 28, no.1, pp. 10–17, Jan. 1993, p.10–17.

[Mead80] C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.

[Sedra87] A. Sedra and K. Smith, *MicroElectronic Circuits*, Holt, Rinehart and Winston, 1987.

[Swanson72] R. Swanson and J. Meindl, "Ion-Implanted Complementary CMOS transistors in Low-Voltage Circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-7, no. 2, pp.146–152, April 1972.

[Sylvester98] D. Sylvester and K. Keutzer, "Getting to the Bottom of Deep Submicron," *Proceedings ICCAD Conference*, pp. 203, San Jose, November 1998.

[Veendrick84] H. Veendrick, "Short-Circuit Dissipation of Static CMOS Circuitry and its Impact on the Design of Buffer Circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-19, no. 4, pp. 468–473, 1984.

习题

提示：请访问 **http://bwrc.eecs.berkeley.edu/IcBook** 以得到最新的思考题、设计题和习题集。通过以电子版而不是印刷版的形式提供这些练习，可以提供一个动态的环境以跟踪当今数字集成电路设计技术的快速发展。